Module as a functional and evolutionary unit in transcription factors and DNA manipulating enzymes

（転写因子およびDNA重合・修復酵素におけるモジュールの機能と進化）

Kei Yura

由良 敬

主論文

Module as a functional and evolutionary unit in transcription factors and DNA manipulation enzymes

「転写因子およびDNA重合・修復酵素における モジュールの機能と分配」

Kei Yura
Group of Computational Structural Biology
Laboratory of Molecular Information and Cellular Regulation
Division of Biological Science, Graduate School of Science, Nagoya University

由良　敬
名古屋大学大学院理学研究科生命理学専攻
機能調節学講座
情報構造生物学

Contents

Abbreviations

1BIA	*bio* repressor
1BPY	DNA polymerase β
1CRO	λ Cro
1LCD	*lac* repressor
1LRD	λ repressor
1OCT	Oct-1 POU specific domain
1PAR	Arc repressor
2CRO	434 Cro protein
2OR1	434 repressor
2TCT	Tet repressor
2WRP	*trp* repressor
3CRO	434 Cro protein
3D	three-dimensional
3FIS	factor for inversion stimulation
3GAP	catabolite gene activator
AAC	amino acid change per codon
brHTH	base-recognition helix-turn-helix
CO	backbone carbonyl oxygen
HhH	helix-hairpin-helix
HKD motif	His-Lys-Asp conserved motif
HTH	helix-turn-helix
NH	backbone amino hydrogen
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
pbHTH	phosphate-binding helix-turn-helix
RHH	ribbon-helix-helix
RMSD	root mean square deviation
sfHTH	scaffold helix-turn-helix
TM	transmembrane

Abstract

Determination of three-dimensional structures of proteins revealed that a protein consists of globular domains. A protein is likely to be evolved by combination of domains, as introns are often found at domain boundaries. However, there are lots of introns inside domains and no clear explanation for the introns has been presented so far. A globular domain can be further divided into several modules, each of which is a contiguous segment with a compact structure composed of about 15 amino acid residues. Module boundaries and intron positions on a corresponding gene are shown to have statistically significant correlation. The correlation suggests that a module was once encoded by a single exon and that introns facilitate forming a different combination of modules by exon shuffling. If this scheme of protein evolution is the case, a module is expected to be a unit of function and a similar module would be found in different proteins. To find a vestige of module shuffling, module organizations of transcription factors and DNA manipulation enzymes were analyzed. A precise assignment of functional sites of these proteins, namely mapping DNA-binding sites, is now possible, because numerous DNA-protein complex structures have been solved by X-ray crystallography and nuclear magnetic resonance spectroscopy measurements.

Decomposition of transcription factors and DNA manipulation enzymes into modules and locating DNA base-binding and phosphate-binding sites showed that there exist three functional types in modules, namely phosphate-binding, base-recognition and scaffold modules. A phosphate-binding module contacts DNA phosphate-backbones only and does not contact DNA bases. A base-recognition module contacts DNA bases by inserting itself into a groove of DNA. A scaffold module does not contact DNA at all. A burden of DNA-

binding is divided by the three types of modules.

Comparison of base-recognition modules and phosphate-binding modules resulted in finding structurally similar modules in different proteins. Prokaryotic transcription factors including catabolite gene activator, 434 Cro protein, 434 repressor, λ repressor, *trp* repressor, factor for inversion stimulation protein, *lac* repressor, *bio* repressor and Tet repressor have a structurally similar base-recognition module that almost corresponds to a helix-turn-helix motif (brHTH module). The C-terminal α helix of brHTH module is inserted into a major groove of DNA. Two modules from DNA polymerase β and one module each from Oct-1 POU domain, 434 Cro protein and Arc repressor have the similar structure and bind a DNA phosphate backbone in a similar manner. An amido hydrogen at the N-terminus of the C-terminal α helix forms hydrogen bonds to phosphodiester oxygen of a DNA backbone. Unexpectedly, the three-dimensional (3D) structures of those phosphate-binding modules resemble that of brHTH module. Therefore the phosphate-binding module was named pbHTH module. In 434 Cro protein, 434 repressor and λ repressor, a scaffold module also had the 3D structure similar to brHTH module, hence named sfHTH module. A similar module is indeed found in different proteins, and also in the same proteins multiple times.

To find a vestige of module shuffling in the whole genome of a single organism, 3D-keynote, an amino-acid-sequence pattern, of pbHTH module was deduced from the requirements on the 3D structure of the module and interactions between DNA and the module. By scanning a total genome of *Synechosystis*, an ORF was found that has the pattern in duplicate, suggesting that the encoded protein binds DNA. With analyses of domain organization and transmembrane regions, the protein was predicted to be located at the periplasmic region and to bind exogenous DNA for natural transformation of *Synechosystis*.

Chapter I.
Overview
-Module as a Unit of Function and Evolution-

I-1. Module organization of globular proteins

Eukaryotic genes are split by introns (Berget *et al.*, 1977; Chow *et al.*, 1977; Sharp, 1994). After transcribing a gene to RNA, introns are spliced out. Gilbert (1978) launched a hypothesis on the origin of introns in a protein coding gene; an intron is a recombination point that changes the combination of exons, regions of genes to be expressed. Evolution by combination of exon can accelerated the formation of novel proteins than by accumulation of point mutations. Blake (1978) argued that if the combination of different exons were allowed, each exon should encode a folded protein fragment. Combination of independent folding units would be more likely to form a stable protein. Doolittle (1978) gave an evolutionary insight to the split gene organization. A split gene organization is the remnant of an ancient gene. The split gene can reduce errors on messenger RNA caused by unfaithful transcription. The present prokaryotic genomes might lost introns during the cause of evolution to speed up replication (Darnell & Doolittle, 1986). The hypothesis is summarized as follows; an exon encodes a peptide fragment that functions by itself and folds by itself, and introns are ancient and can be lost. The question that should be addressed next is to find a unit that was encoded by an ancient exon.

Globular protein were found to be partitioned into compact sub-structure termed module. A module is determined by 3D structure of a protein, calculating Cα-Cα distances and dividing the protein into locally compact structures. The average length of module is about 15 residues. Module boundaries were first determined on a distance map, a triangular plot to depict

distances of two Cα atoms. On a residue with the Cα atom close to all the other Cα atoms lies a module boundary (Gō, 1981; Gō, 1983). The method had ambiguity in deciding closeness of Cα atom pairs and was not applicable to big proteins. To eliminate the ambiguity and to enable to determine module organization of big proteins, centripetal and extension profiles were introduced (Gō & Nosaka, 1987). The profiles plot smoothed value of a mean square distance of Cα atoms from *i*th residue within the range of *k* residues. Local minima of the centripetal profile and local maxima of the extension profile are candidates of module boundary. Judging the stability of the candidate boundaries against several *k*s, module boundaries were finally determined. The procedure is now fully automatized.

Module boundaries and positions of intron are shown to correlate in many proteins. On hemoglobin, the protein on which module structure was first discovered, a position of one intron was predicted from the modular structure and later an intron at the place was found on leghemoglobin (Gō, 1981). Positions of introns of lysozyme (Gō, 1983), cytochrome c (Gō, 1985), ovomucoid third domain (Gō, 1985) and triose phosphate isomerase (Gō & Nosaka, 1987; Gilbert & Glynias, 1993) were correlated with module boundaries. Intron positions on triose phosphate isomerase was predicted by somewhat different method of module boundary determination by Gilbert *et al* (1986). An intron was found at one of the predicted locations on the gene from mosquito *Culex tarsalis* (Tittiger *et al.*, 1993). Statistical significance of correlation was tested on enzymes of glycolysis (Gō & Noguti, 1995) and well conserved proteins (de Souza *et al.*, 1996). The correlation suggests that a module be once encoded by a single exon. Module shuffling driven by exon shuffling (Gilbert, 1978) might create a new protein. The correlation suggests that module may be a unit of evolution (Gō, 1985).

Module is also a unit of function, as it was shown experimentally by isolating a single module or by swapping modules in a native protein. Some isolated modules of barnase, a bacterial RNase, had an enzymatic activity (Yanagawa *et al.*, 1993). Module F4 of hemoglobin α subunit was exchanged to module F4 of β subunit and the chimera subunit had the α-like heme environment with β-like subunit interface; the chimera subunit preferred to interact with α subunit (Wakasugi *et al.*, 1994; Inaba *et al.*, 1998). A single module in isocitrate dehydrogenase that binds NADP was exchanged to a corresponding module in NAD-specific isopropylmalate dehydrogenase and the coenzyme specificity was exchanged without loss of activity (Yaoi *et al.*, 1996).

A module is considered to be a unit of evolution from the correspondence of gene organization. Module shuffling might created a novel protein. The scenario suggests that remnants of module shuffling be found on the present proteins. The best target in search of the remnants is proteins that interact with DNA. An assignment of functional sites on those proteins is easy due to determination of protein-DNA complex structures by X-ray and nuclear magnetic resonance spectroscopy measurements.

I-2. Proteins that control and manipulate DNA

Transcription factors and DNA manipulation enzymes such as restriction, modification, replication and repair enzymes, and nucleotide polymerases play central roles in organisms. Transcription factors turn gene transcriptions on and off. The proteins bind DNA with base specificity. Direct interactions between DNA bases and protein side chains facilitate DNA sequence specificity (Pabo & Sauer, 1992). DNA with bound transcription factors is often bent, which promotes or inhibits RNA polymerase to bind the DNA. Some transcription

factors directly bind to a promoter region and inhibit RNA polymerase from binding to the promoter. Alternatively, by binding to remote regions from a promoter, the factor starts to recruit other transcription factors onto DNA. Repair enzymes find damaged bases by checking bases directly, or by checking phosphate backbone conformations that are deformed because of the damage. Specific interactions with damaged bases initiate cleavage or exchange reactions (Myers & Verdine, 1994; Tainer *et al.*, 1995; Vassylyev & Morikawa, 1997). Nucleotide polymerases copy the template DNA to new nucleic acid sequences. The enzymes mainly interact nonspecifically with DNA; the interactions are mainly between the enzymes and DNA backbones (Joyce & Steitz, 1994).

Some of the transcription factors and DNA manipulation enzymes are ancient proteins, because those proteins are common to most of the organisms existing on the earth today; the common existence strongly suggests that a common ancestor of the organisms had those proteins. The proteins must have evolved at the time that DNA had started to store genetic information. Present diversity of transcription factors is mostly based on different combinations of domains (Zuckerkandl, 1994). Homeodomains are found in proteins that control transcription of gene encoding proteins for differentiation of cells in higher organisms. Differentiation is known to be specific for eukaryotes, though three-dimensional (3D) structure of homeodomain is surprisingly similar to a prokaryotic repressor (Assa-Munt *et al.*, 1993). A recent study on of homeodomain-like proteins in plant plastid also suggests that prototype of homeodomain might exist before the divergence of eubacteria and eukaryotes (Yura & Gō, 1997). In eukaryotes, homeodomains are integrated into big proteins. The proteins exist as a multigene family on chromosomes (Gehring *et al.*, 1994).

The fundamental function of transcription factors and DNA manipulation

enzymes is to bind DNA. The mechanism to attain this function can be elucidated by solving DNA-protein complex structures. Advances in X-ray crystallographic and nuclear magnetic resonance measurements enabled to disclose the DNA-protein complex structures. Up to 1997, 51 DNA-protein complex structures were deposited to Protein Data Bank (PDB) (Bernstein *et al.*, 1977) as shown in Table I-1. Collection of DNA-protein complex structures first revealed that transcription factors and DNA manipulation enzymes have domain structures; a DNA-binding domain and domains to interact with protein and/or effectors. Transcription factors and DNA manipulation enzymes tend to form oligomer to cover wide range of DNA sequences (Harrison, 1991). Zif268 protein, a rapidly activated transcription factor after stimulation in mouse fibroblast cell (Christy *et al.*, 1988), has similar three DNA-binding domains in tandem (Pavletich & Pabo, 1991). TATA-box binding protein (TBP) has apparent internal duplication (Kim *et al.*, 1993; Kim *et al.*, 1993). Prokaryotic repressors including 434 repressor (Aggarwal *et al.* 1988) and restriction enzymes such as Eco RV (Kostrewa & Winkler, 1995) binds DNA in dimer (Figure 1-1).

Residues that directly interact with DNA bases are often located on a similar local structure in DNA-binding domains of transcription factors and DNA manipulation enzymes. The local structures were named after their secondary structures as helix-turn-helix (HTH) or helix-loop-helix (HLH) motifs or after their characteristic ligands or residues as zinc finger or leucine zipper motifs (Luisi, 1995) (Table I-1). A motif first found in DNA-binding domain was HTH motif (Harrison & Aggarwal, 1990). A number of transcription factors and DNA manipulation enzymes with the HTH motif surpasses that of the other DNA-protein complexes in PDB. In the total genome of cyanobacteria, an HTH protein was predicted to be the fourth largest group

Table I-1: DNA-protein complex structures in PDB

Name	PDB ID	reference
Helix-Turn-Helix Protein		
434 repressor	2OR1	Aggarwal *et al.*, 1988
λ repressor	1LRD	Jordan & Pabo, 1988
trp repressor	1TRO	Otwinowski *et al.*, 1988
λ Cro	4CRO	Brennan *et al.*, 1990
434 Cro protein	3CRO	Mondragon & Harrison ,1991
catabolite gene activator	1CGP	Schultz *et al.*, 1991
purine repressor	1PNR	Schumacher *et al.*, 1994
lactose operon repressor	1LBG	Lewis *et al.*, 1996
engrailed homeodomain	1HDD	Kissinger *et al.*, 1990
antennapedia protein	1AHD	Billeter *et al.*, 1993
MAT alpha2 homeodomain	1APL	Wolberger *et al.*, 1991
MAT a-1	1YRN	Li *et al.*, 1995
Oct-1	1OCT	Klemm *et al.*, 1994
paired domain	1pdn	Xu *et al.*, 1995
c-Myb	1mse	Ogata *et al.*, 1994
ETS1	2stw	Werner *et al.*, 1995a
Zinc Finger Protein		
Zif268	1ZAA	Pavletich & Pabo, 1991
tramtrack protein	2DRP	Fairall *et al.*, 1993
YY1	1UBD	Houbaviy *et al.*, 1996
glucocorticoid receptor	1GLU	Luisi *et al.*, 1991
estrogen receptor	1HCQ	Schwabe *et al.*, 1993
retinoic acid receptor	2NLL	Rastinejad *et al.*, 1995
GATA-1	1GAT	Omichinski *et al.*, 1993
pyrimidine pathway regulator 1	1PYI	Marmorstein & Harrison, 1994
GAL4	1D66	Marmorstein *et al.*, 1992
Leucine Zipper Proteins		
c-Jun proto-oncogene	1FOS	Glover & Harrison, 1995
GCN4	1DGC	Ellenberger *et al.*, 1992
MyoD	1MDY	Ma *et al.*, 1994
Ribbon-Helix-Helix Protein		
met repressor	1CMA	Rafferty *et al.*, 1989
arc repressor	1PAR	Raumann *et al.*, 1994

Table I-1: (Continued)

Name	PDB ID	reference
High Mobility Group Protein		
sry	1HRY	Werner *et al.*, 1995b
LEF-1	1LEF	Love *et al.*, 1995
MADS Domain Protein		
serum response factor	1SRS	Pellegrini *et al.*, 1995
HU/INF Proteins		
integration host factor	1IHF	Rice *et al.*, 1996
Others		
transcription factor IIA	1YTF	Tan *et al.*, 1996
transcription factor IIB	1VOL	Nikolov *et al.*, 1995
TATA-box binding protein	1YTB	Kim *et al.*, 1993; Kim *et al.*, 1993
nuclear factor κ-B	1NFK	Ghosh *et al.*, 1995
bovine papillomavirus-1 E2	2BOP	Hegde *et al.*, 1992
tumor suppressor p53	1TUP	Cho *et al.*, 1994
Enzymes		
DNase I	2DNJ	Lahm & Suck, 1991
DNA polymerase	1KLN	Beese *et al.*, 1993
DNA polymerase β	1BPY	Sawaya *et al.*, 1994
Hin recombinase	1HCR	Feng *et al.*, 1994
PuvII endonuclease	1PVI	Cheng *et al.*, 1994
Eco RI	1ERI	McClarin *et al.*, 1986
Eco RV endonuclease	1RVA	Kostrewa & Winkler ,1995
HhaI methyltransferase	1MHT	Klimasauskas *et al.*, 1994
replication terminator protein	1ECR	Kamada *et al.*, 1996
endonuclease V	1VAS	Vassylyev *et al.*, 1995
FokI restriction endonuclease	1FOK	Wah *et al.*, 1997

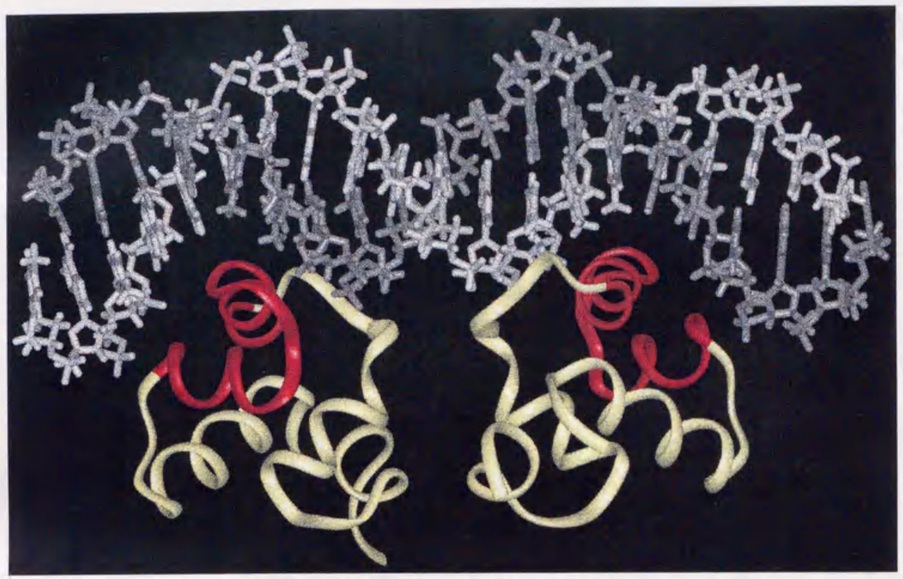

Fig I-1. 3D structure of 434 Cro protein dimer of 434 bacteriophage bound to its operator DNA depicted by backbone tube model. Red parts are the helix-turn-helix motifs. The motif is not the only segment that binds DNA, but not all part of the protein bind DNA. The coordinate was from Protein Databank under the ID of 3CRO (Mondragon & Harrison, 1991).

(Koonin & Galperin, 1997).

The classification of DNA-binding domains of transcription factors and DNA manipulation enzymes has been focused on DNA-binding motifs and has displayed a variety in base recognition sites, namely various motifs, and common mechanism in DNA binding, namely dimerization (Harrison, 1991). The classification, however, has ignored the fact that DNA-binding domains do not consist only of the motif, but built also with other structures (Suzuki & Brenner, 1995). Stress has been put on motifs that recognizes base sequences, but DNA-binding domains do not interact only with DNA bases. Interactions with DNA phosphate backbones and proteins were noted to play an important role in determining base sequence specificity (Pabo & Sauer, 1992). In addition to that, every residue in DNA-binding domains obviously does not contact DNA bases and backbones (Figure I-1).

I-3. Module as a functional unit in transcription factors and DNA manipulation enzymes

DNA-binding domains of transcription factors and DNA manipulation enzymes are decomposable to modules as shown in Figure I-2. When base contact residues and phosphate (ribose) contact residues are mapped on modules, interaction sites of bases and phosphates are found to be localized to some modules. Those modules can be further classified into two; a module that is only in contact with phosphate (ribose) and a module that contacts bases.

HTH motif, one of the well-known motifs, almost corresponds to a single module (Figures I-1 & I-2). HTH motif is composed of two helices connected by a turn with the angle of about 120 degrees (Harrison, 1991). HTH motif in 434 Cro protein of phage repressor, module M3, is almost equivalent to the motif (Yura *et al.*, 1993). As side chains of residues on the C-terminal α helix form specific hydrogen bonds to DNA bases, the module is named base-recognition (br) HTH module. The module is situated in the major groove of DNA (Figure I-3). The structural details of brHTH module are elucidated in Chapter II.

DNA-protein interactions are not accomplished by base-recognition modules alone. Interactions between DNA phosphate backbones and the proteins indirectly facilitate sequence specific binding of the proteins (Pabo & Sauer, 1992). When DNA-binding domains are decomposed into modules, modules that contact only to DNA phosphate backbones were found (Figure I-2). In 434 Cro protein, modules M2 and M4 contact only to DNA phosphate backbones (Figures I-2). Of them, module M2 assumes a 3D-structure surprisingly similar to brHTH module, hence termed phosphate-binding (pb) HTH module (Yura *et al.*, 1999a). The pbHTH module forms hydrogen bonds between a DNA phosphodiester oxygen and a protein backbone amino hydrogen

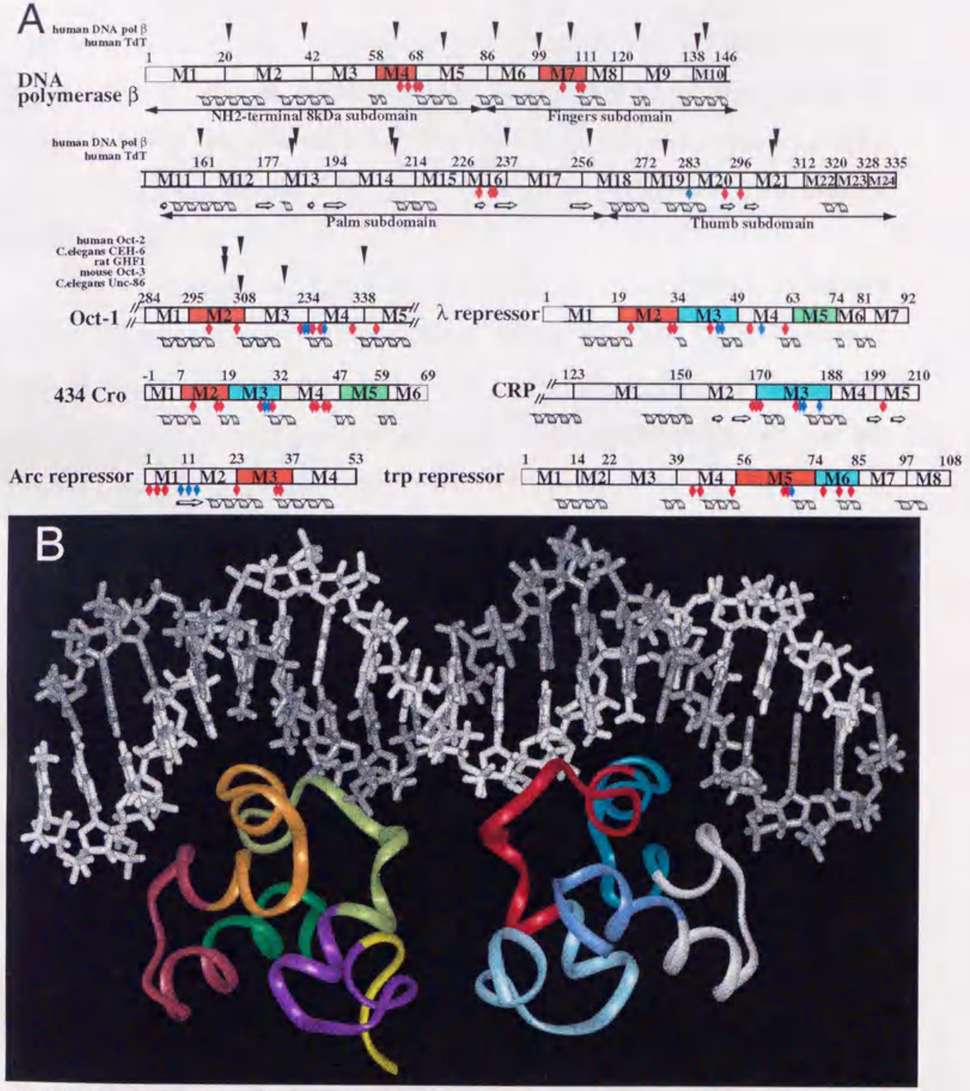

Fig I-2. (A) Module organization of a DNA repair enzyme and six transcription factors. A blue box indicates base-recognition helix-turn-helix module, an orange box is phosphate-binding helix-turn-helix module and a green box is scaffold helix-turn-helix module. A red diamond mark on module diagrams shows a DNA phosphate/ribose contact residue and a blue mark a base contact residue. Positions of secondary structures are depicted by coils (α helices) and arrows (β strands). Intron positions on genes of DNA polymerase β and its homolog, and Oct-1 and its homologs are shown by a wedge. Detail of intron-module correspondence is discussed in Chapter III. Note that the phosphate-binding and the base-recognition modules of trp repressor have special features. The phosphate-binding module M5 contacts a DNA base and the base-recognition module M6 does not contact a DNA base. They are assumed to be evolutionary intermediates. (B) 3D structure of 434 Cro protein coloured in module. A green module at the far side and steel blue module at the near side is an N-terminal module of each subunit. Note that the helix-turn-helix motif in Fig I-1 almost corresponds to a single module.

at the N-terminus of the C-terminal α helix, whereas brHTH module forms hydrogen bonds between DNA bases and side chains of residues on the C-terminal α helix. The pbHTH module of 434 Cro protein is located on a DNA phosphate backbone (Figure I-2C). The structural details of pbHTH module are elucidated in Chapter III.

Not all modules of DNA-binding domains contact DNA molecules (Figure I-2B). There are modules that evidently facilitate base-recognition modules and phosphate-binding modules by placing them to a proper location to bind DNA; the module is termed scaffold modules. Modules M1, M5 and M6 of 434 Cro protein are scaffold modules. Of them, module M5 has 3D structure, again, similar to HTH module and hence termed scaffold (sf) HTH module (Yura & Gō, 1995).

Transcription factors and DNA manipulation enzymes interact with DNA using two different types of modules; base-recognition and phosphate-binding modules (Figure I-2). The base-recognition module is inserted into a groove of DNA and form hydrogen bonds with DNA bases, hence achieves base specific interactions. The phosphate-binding module is located on a backbone of DNA and contact only to the backbone (Figure I-3). Two types of modules generate a façade to bind DNA in the domain. Two kinds of labour, namely base contact and phosphate contact are divided by the two types of module. Scaffold modules are located on the other side of the façade, evidently to sustain the two types of DNA-binding modules in appropriate positions. The DNA-binding domains of transcription factors are divided into three types of modules without extra regions (Figure I-2).

It is intriguing to find that a size of module is adjusted to a concave and a convex of DNA structure. A base-recognition module is situated in a major groove with a handle protruding to scaffold modules (Figure I-3B). A whole

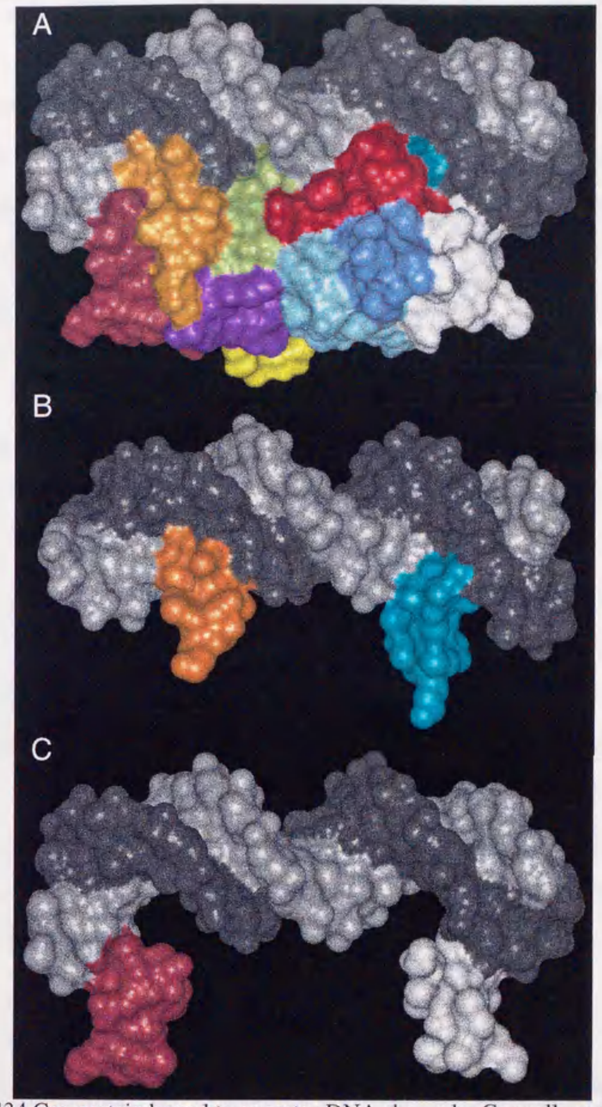

Fig I-3. (A) 434 Cro protein bound to operator DNA shown by Connolly model (Connolly, 1983). The model represents a solvent accessible surface of the protein-DNA complex. The molecule is viewed from the same side of Fig I-2(B). The protein is coloured in module and DNA is coloured in dark and light gray. The model indicates that three modules in each subunit, namely modules in white, deep blue and red in one subunit and those in green, orange and brown in the other subunit interact with DNA. Modules in steel blue and sky blue in one subunit and those in yellow and purple in the other subunit do not interact with DNA. Compare Fig I-2(B) to identify subunits. (B) Interactions of base-recognition helix-turn-heilx modules M3 in both subunits of 434 Cro protein bound to DNA. Connolly surface was calculated after eliminating other modules. The modules are penetrated into DNA major grooves. The modules scarcely contact DNA minor groove and backbones. The size of the modules well-fit to the width of a major groove. The handles protruding out in the air is buried into the core in the intact protein. (C) Interactions of phosphate-binding helix-turn-helix modules M2 in both subunits of 434 Cro protein bound to DNA. The modules are placed on DNA backbones. The modules scarecly contact DNA grooves and bases.

part of the domain is too large to be located in the groove. Similarly, a phosphate-binding module covers a part of a DNA backbone without extra contact to other parts of DNA (Figure I-3C). Length of modules in globular proteins are about 15 residues on average (Figure I-2A). This length coincides with a length of polypeptide that is estimated to form a nucleus of folding in a protein (Wetlaufer, 1973). The average length of module also corresponds to a length of ribonucleotide oligomer that can be formed under template-directed reactions with metal cations (Joyce, 1987). The size of module, probably limited by the ability to fold by itself and by the length of an RNA template, coincided with the size of a DNA groove and hence modules were conveniently used to interact with DNA.

Protein-DNA interactions are established in the early stage of molecular evolution, since most of the organisms have the similar transcription factors and DNA manipulation enzymes. In purine repressor, a segment that assumes an α helix when bound to DNA was found to be a random coil in DNA-free form (Nagadoi *et al.*, 1995). Stabilization of 3D structure by DNA interaction suggests that the interactions outdate evolution of the DNA-binding domain. A scenario that an established domain structure turns to destabilized itself by initiating interactions with DNA is hard to come about.

I-4. **Repetitive appearance of the similar module in different proteins**

The brHTH module that is almost equivalent to HTH motif was found in phage repressors (Yura *et al.*, 1993). Their 3D structures as well as amino acid and nucleotide sequences were significantly similar (Ohlendorf *et al.*, 1983; Yura *et al.*, 1993). The pbHTH module that interacts with DNA phosphates with the similar manner was found in 434 Cro protein, Arc repressor, POU-

specific domain and DNA polymerase β (Yura *et al.*, 1999a). The sfHTH module that does not bind DNA, but serves as a foundation for other modules was found in 434 Cro protein and λ repressor (Yura *et al.*, 1993). This sfHTH modules were also found in proteins that apparently do not bind DNA (Yura & Gō, 1995).

The distribution of each type of HTH modules in different proteins supports the scenario that a module with similar 3D structure was used in different module combination to evolve different proteins. Similarity in 3D structures of the modules as well as similarity in amino acid sequences within each type of HTH modules seem to be remnants of a common ancestor (Yura *et al.*, 1993). Structural and functional details in similarity of brHTH modules and pbHTH modules are extended in Chapters II and III, respectively. Structural similarity in all types of HTH modules could lead to an inference that the three types of HTH modules had a common ancestor. A functional intermediate between the brHTH module and the pbHTH module was found in *trp* repressor; the repressor uses a pbHTH-like module to specifically interact with DNA bases (Yura *et al.*, 1999a). Marginal sequence similarity in the three types of HTH modules was noticed (Yura *et al.*, 1993). An extent of distribution of HTH module remains to be shown, yet preliminary comparison of all module structures of the proteins with the known 3D structures resulted in finding more HTH modules, indicating that the HTH module was widely used.

Which type of function was loaded on HTH module first? Out of protein, RNA and DNA, RNA is considered to be the first to appear (Joyce & Orgel, 1993). A scenario of molecular evolution tells that a short peptide started to interact with RNA enzymes as a scaffold (Jay & Gilbert, 1987). As a phosphorous atom is a fundamental component of RNA, finding phosphorous atoms is considered to be the most effective way to bind RNA. The scenario,

then, suggests that the first HTH module bound a phosphate backbone of RNA.

Success in recruiting an HTH module into different combination of modules requires that the HTH module maintains its 3D structure in different context of modules. Ikura *et al.* (1993) showed by NMR measurements that an isolated module had a tendency to take a 3D structure similar to its structure in an intact enzyme. Conformation of a module is mostly determined by intra-module interactions (Noguti *et al.*, 1993). Mechanical stability of a single module was shown by *in vacuo* and solution computer simulations (Takahashi *et al.*, 1997).

The appearance of HTH modules in different proteins was likely to be the result of module shuffling (Gō, 1995) driven by exon shuffling (Gilbert 1978). The correspondence of a single module and a single exon is suggested in correspondence of module boundaries and intron positions as explained in II-2. The HTH modules were often found in transcription factors of eubacteria that have no introns in DNA for mRNA coding sequences. Prokaryotic genomes were suggested to have introns to overcome unfaithful transcription at first (Doolittle, 1978), and to have lost introns to speed the replication after faithful transcription had achieved (Darnell & Doolittle, 1986).

1-5 Repetitive appearance of the similar module in the same proteins

Distribution of similar module is not limited to different proteins, but was found in single transcription factors and DNA manipulation enzymes. Phage repressor, 434 Cro protein, has three HTH modules, λ repressor has two and DNA polymerase β two (Yura *et al.*, 1993; Yura *et al.*, 1999a). ComEA of cyanobacteria, a putative protein for natural transformation that binds exogenous DNA, was predicted to have two pbHTH modules (Yura *et al.*,

1999b) A preliminary survey of protein 3D structure database showed that TFIIB and endonuclease III also had multiple HTH modules within them.

The duplication of HTH modules in those proteins suggests a new model of protein evolution; duplication of modules. Duplication of domains is a well-established protein evolution model. Fibronectin is one of the well-cited examples that has three types of multiple domains each separated by introns. Homologous and non-homologous recombination can drive duplication (Patthy, 1991; Doolittle, 1995; Patthy, 1996). McLachlan (1987) exemplified numerous cases of domain multiplication observed through 3D structure comparisons.

Contrary to the model of protein evolution by domain duplication, a few models of domain evolution have been presented. Eck and Dayhoff (1966) first discovered a four-residue sequence repeat within ferredoxin. *Clostridum pasteurianum* ferredoxin has variation of ADSG sequences 13 times. Repeat sequences were also found in cytochome and hemoglobin (Canter & Jukes, 1966), subtilisins BPN' and Carlsberg (Smith *et al.*, 1968) and histone H1 (Ohno, 1987). Most of the repeats were, however, shown to be statistically insignificant and the theory has lost statistical justification (McLachlan, 1972), except for a few obvious cases (McLachlan, 1987). Statistical measurement of amino acid sequences of a domain reached to claim that the pattern of amino acid sequences is nothing but random (White, 1994).

Comparison of 3D structures of homologous proteins revealed that proteins with weak identity retained structural similarity in protein core (Chothia & Lesk, 1986). Evolutionary relationship of proteins remains in their 3D structures much longer than in their amino acid sequences (Doolittle, 1995). Bajaj & Blundell (1984) found that some of the motifs were arranged symmetrically in the protein 3D structure and suggested internal duplication of a segment within a domain. Their model was, however, limited to a case with an

apparent sequence similarity.

Repetitive appearance of modules detected by their 3D structure similarities suggests an evolution of protein by module duplication. Approximately 15 amino acids were randomly joined and that sequences with adequate property started to duplicate to form a domain. The necessary property of the peptide was a tendency to take a 3D structure in itself (Ikura *et al.*, 1993; Noguti *et al.*, 1993; Takahashi *et al.*, 1997) and have a function, probably that to interact with phosphate.

The module duplication evolution model is based on 3D-structure analyses of transcription factors and DNA manipulation enzymes, so that the generality of the model needs to be elucidated. In addition to that, the model addresses a fundamental question that how many modules there were to build the present proteins. The present exon distribution suggested that 1,000 to 7,000 exons were needed to construct all the present proteins (Dorit *et al.*, 1990; Dorit *et al.*, 1991). A number of ancient modules must be smaller, since a single exon encodes multiple modules (Figure I-2A).

Chapter II.
Correspondence of a helix-turn-helix motif to a single module and repeat of the module in transcription factors

Abstract: Some prokaryotic repressors consist of several α-helices connected with turns. We report here that these repressors are decomposable into helix-turn-helix modules and their connectors. A module is defined as a compact structural unit with consecutive amino acid residues in a globular protein. Helix-turn-helix motif is one of the common motifs observed in DNA-binding proteins. The motif interacts with DNA double helix and recognizes specific base sequences. It is assumed that the helix-turn-helix motif appears only once in ten prokaryotic transcriptional repressors of which 3D structures have been determined by X-ray crystallographic or nuclear magnetic resonance measurements. Each of the helix-turn-helix motifs in the ten proteins corresponds approximately to a single helix-turn-helix module consisting of approximately 13 amino acids. Identification of modules of ten prokaryotic repressors and comparisons of their tertiary structures led to the conclusion that three of these DNA-binding proteins contain more than one helix-turn-helix module with a structure similar to the helix-turn-helix motif. The difference in DNA base recognition ability in these helix-turn-helix modules is ascribed to a difference in size of a side chain at the fifth residue from Gly, on the turn. The difference in module organization of these DNA-binding proteins paves the way for further classification of the DNA-binding proteins with the helix-turn-helix motif. The structural repertoire of these transcriptional regulators was increased through different utilizations in the number of helix-turn-helix and other modules.

II-1. Helix-turn-helix motif and helix-turn-helix module

The determination of 3D structures of DNA-binding proteins has revealed several common structures for protein-DNA interactions (Schleif, 1988) . Many

DNA-specific binding proteins interact with DNA by inserting a recognition helix in the major groove of DNA (Harrison, 1991). In the helix-turn-helix (HTH) motif, named after its secondary structure (Brennan & Matthews, 1989; Steitz, 1990), two helices connected with a short bend form an angle of 120 degrees. The COOH-terminal α helix is a recognition helix. The appearance of the HTH motif in a wide range of species is striking; it has been found in prokaryotic repressors and in the eukaryotic homeodomain by X-ray crystallographic analysis (Kissinger *et al.*, 1990) and 2-D NMR study (Otting *et al.*, 1990).

The amino acid sequences of the HTH motifs were also compared. It was suggested that the HTH motifs had derived from a common precursor and the ability of DNA binding for specific operators evolved by amino acid replacements within the HTH motif (Ohlendorf *et al.*, 1983). Structures corresponding to the HTH motif have not been found in other proteins (Ohlendorf *et al.*, 1983; Takeda *et al.*, 1983; Brennan & Matthews, 1989), except for a portion in HU protein (Tanaka *et al.*, 1984) , in cytochrome c peroxidase (Richardson & Richardson, 1988), in L11 ribosomal protein (Hinck *et al.*, 1997) and in ribosomal L7/L12 protein (Richardson & Richardson, 1988; Rice & Steitz, 1989). Rice and Steitz (1989) further argued that the protein might interact with RNA. Li *et al.* (1992), however, disproved the nucleic acids binding ability of tobacco chloroplast ribosomal protein L12.

The prokaryotic repressors with the motif consist of several α-helices connected with bends and of a few short β-strands in a few cases (Anderson *et al.*, 1981; Schultz *et al.*, 1991). Since the α-helices and bends exist alternatively along the sequence, these conformations could be decomposed into topologically repetitive units with two helices and a bend in between by dividing the polypeptide segments on the helices.

Modules are defined in globular proteins as compact polypeptide segments that are separated in space from one another (Gō, 1981). Correspondence of module boundaries with the position of introns in eukaryotic genes has been detected in haemoglobin (Gō, 1981), lysozyme (Gō, 1983), cytochrome *c* (Gō, 1985), ovomucoid third domain (Gō, 1985) and triose phosphate isomerase (Gilbert *et al.*, 1986; Gō & Nosaka, 1987). This and the exon-shuffling hypothesis (Gilbert, 1978) leads to the assumption that the module is a unit of protein evolution. A module is detectable in a triangular distance map by partitioning a whole protein into contiguous segments in such a manner that the largest number of pairs of Cα atoms separated less than a certain distance are included in each module (Gō, 1981). The boundaries of modules tend to lie on α-helices or β-strands (Gō & Nosaka, 1987) and a module usually includes one turn. The module supports the idea that the protein is divided on its secondary structures. We describe here our analyses of the architectures and the evolutionary relationship of proteins with the HTH motif, based on module structures.

II-2. Module organization of ten transcription factors

Coordinate of HTH proteins

Ten DNA-binding proteins with the HTH motif were selected from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977); phage 434 Cro protein (2CRO) (Mondragon & Harrison, 1991), phage 434 repressor (2OR1) (Aggarwal *et al.*, 1988), operator binding domain of λ repressor (1LRD) (Jordan & Pabo, 1988; Clarke *et al.*, 1991), λ Cro (1CRO) (Anderson *et al.*, 1981), COOH-terminal domain of catabolite gene activator (3GAP) (Schultz *et al.*,

1981), *trp* repressor (2WRP) (Otwinowski *et al.*, 1988), factor for inversion stimulation (fis) (3FIS) (Yuan *et al.*, 1991), *lac* repressor (1LCD) (Chuprina *et al.*, 1993), *bio* repressor (1BIA) (Wilson *et al.*, 1992) and Tet repressor (2TCT)(Kisker *et al.*, 1995). The structures were all determined by X-ray crystallography except for *lac* repressor the structure of which was determined by 2D-NMR measurement. The resolutions of the proteins determined by X-ray crystallography exceeded 2.5 Å.

Module assignment

Modules of the ten proteins are assigned by centripetal and extension profiles (Gō & Nosaka, 1987). Module boundaries are located in the centre of a local segment of a protein. The centripetal profile detects residues at the centre of the local segment by calculating the mean square distance between the Cα atom of the *i*th residue and that of the other Cα atoms within the range of $\pm k$ from the *i*th residue. Therefore, it is formulated as,

$$F_i = \sum_{i-k \le j \le i+k} r^2_{ij} / (2k+1).$$

Several values of k (k = 15, 20, 25, 30, 35, 40 and 45) are taken. In the calculation of F_i, $i-k$ and $i+k$ should not exceed the NH_2- and COOH-terminal of the protein. The local minima of F_i against i are candidates for module boundaries. Compactness of the fragment between two boundaries is confirmed by the extension profile (Gō & Nosaka, 1987). The position of the local minimum of the centripetal profile or the local maximum of the extension profile sometimes depends on the range k. An unstable minimum or maximum merges into a close stable minimum or maximum. A minimum or a maximum, too close to the terminal of the sequence, is neglected even if it is stable. The minimum length of the module is limited to five. The process is all carried out

automatically on a computer (Gō *et al.*, in preparation) .

The result of module assignment is drawn on distance maps (Figure II-1). A pair of Cα atoms further than 18.0Å is marked in gray in a triangular space. The module boundaries assigned by the profiles are drawn on the maps. Triangles on the diagonal lines partitioned by module boundary lines scarcely contain the gray mark; this confirms that the assigned modules are compact. Rectangles partitioned by module boundary lines in the triangular maps do contain the gray mark, thereby indicating that the modules are well separated from one another. With identification we find that 28 out the 52 modules have helices at both termini and a bend or turn in the middle region. The following modules form the HTH motif; 434 Cro protein (2CRO) M3 (20-32), 434 repressor (2OR1) M2 (19-32), λ repressor (1LRD) M3 (35-49), λ Cro protein (1CRO) M3 (20-31), catabolite gene activator (3GAP) M3 (170-187), *trp* repressor (2WRP) M5 (72-85), fis (3FIS) M5 (79-88), *lac* repressor (1LCD) M2 (9-17), *bio* repressor (1BIA) M3 (25-38) and Tet repressor (2TCT) M3 (28-42) (Figure II-2). These were termed DNA base-recognition HTH modules.

II-3. Structure comparison among modules

Three-dimensional structure of the modules was compared by superimposing the Cα atoms of two modules. A root mean square deviation (RMSD) is calculated after least-square fitting of the two by the method of McLachlan (1979). To superimpose two modules, three more residues are added on both termini of one of the modules. This is because of a slight ambiguity at the module boundaries. The shorter module is shifted along the longer one from its NH_2- to COOH-terminus and all the RMSDs of the best fit between them are calculated. The least value amongst all of them is taken as the RMSD between

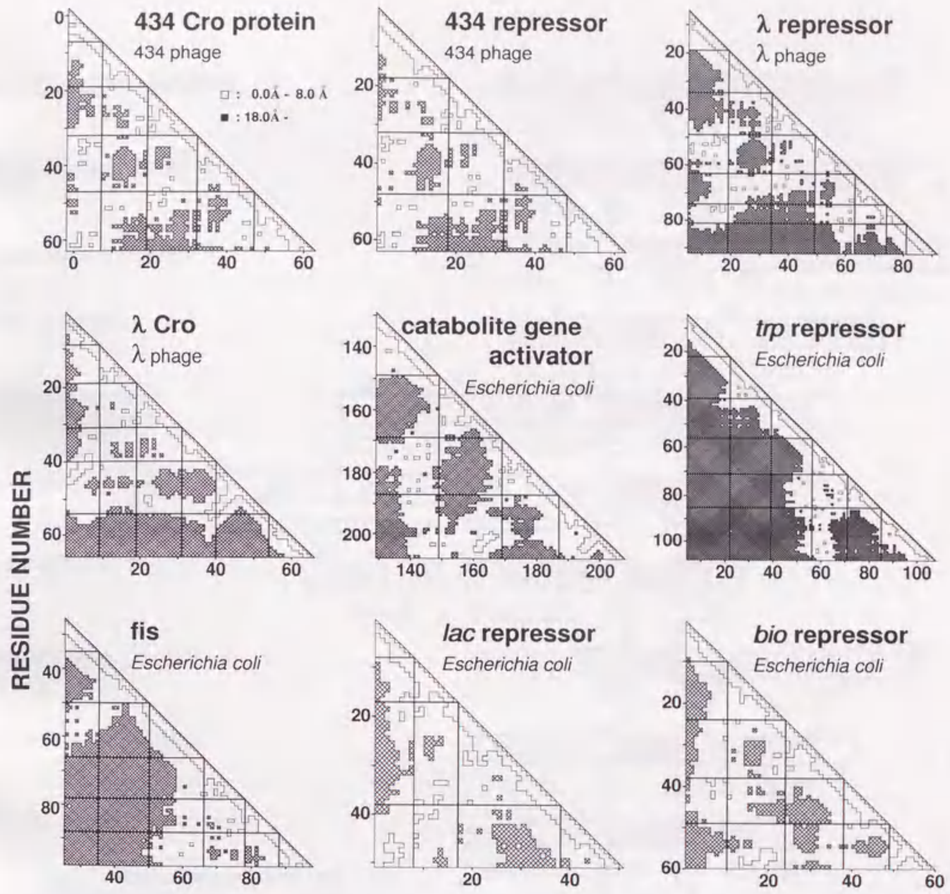

Fig II-1. Distance maps showing module boundaries of phage 434 Cro protein, phage 434 repressor, operator binding domain of λ repressor, λ Cro, COOH-terminal domain of catabolite gene activator, trp repressor, factor for inversion stimulation (fis), NH-terminal domain of lac repressor, NH-terminal domain of bio repressor and NH-terminal domain of Tet repressor. Distances between the ith and jth Cα atoms are plotted in a triangular space, in which both ordinate and abscissa are the residue number. The shaded region represents pairs of Cα atoms separated more than 18A. The region surrounded by lines represents the pairs of Cα atoms closer than 8A. Module boundaries drawn by horizontal and vertical lines in a triangular space are determined by centripetal and extension profiles (Go & Nosaka, 1987). The residue number of 434 Cro protein starts with -1 and that of trp repressor and Tet repressor with 2. The structure of the following regions was not determined by X-ray crystallography; 64-69 of 434 Cro protein, 64-69 of 434 repressor, 1-5 of λ repressor, 209 of catabolite gene activator, 2-4 of trp repressor and 1-25 of fis.

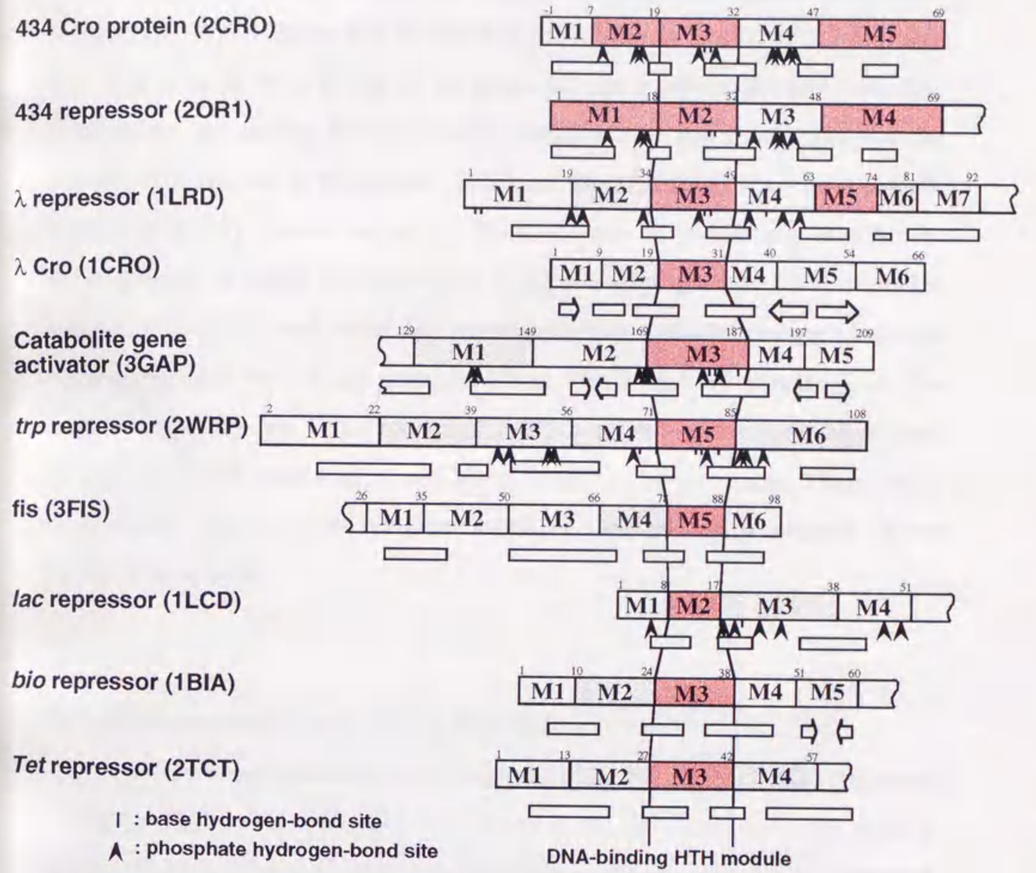

Fig II-2. Module organizations and secondary structures. Boxes with symbols represent modules and the numbers on the top of the boxes indicate the residue number of the module boundaries. A rectangle below the boxes indicates an α-helix, a small rectangle indicates a 3_{10} helix and an arrow indicates a β-strand. Red modules have similar structures. Dotted modules have only topologically the same structure with 434 Cro protein module M3. Modules drawn at the centre of the picture consist of the HTH motif. A wedge below the boxes means a site which forms a hydrogen bond with a phosphate of DNA backbone. A thin bar below the boxes means a site which forms a hydrogen bond with a base. It recognizes a specific sequence of DNA. Each site was determined in the article; 434 Cro protein is by Mondragon and Harrison (1991), 434 repressor by Aggarwal *et al.* (1988), λ repressor by Jordan & Pabo (1988) and Clarke *et al.* (1991), catabolite gene activator by Schultz *et al.* (1991) and *trp* repressor by Otwinowski *et al.* (1988). *lac* repressor by Chuprina *et al.* (1993). High resolution crystallographic analysis of λ Cro-DNA, fis-DNA, bio repressor-DNA and Tet repressor DNA complexes have not been reported. The structure of the undetermined regions is included in preceding or following modules, namely 64-69 of 434 Cro protein into module M5, 64-69 of 434 repressor into module M4, 1-5 of λ repressor into module M1, 209 of catabolite gene activator into module M5 and 2-4 of *trp* repressor into module M1.

the two modules.

Fifteen of all the modules were superimposed against the DNA recognition HTH modules 434 Cro protein M3 within an RMSD of 1.3 Å (Table II-1, Figure II-3). Five of the 15 modules did not compose the DNA-binding HTH motif. According to the structural analysis of the significance of the RMSD (Remington & Matthews, 1980), an RMSD of less than 1.5Å for 13-residue randomly chosen sequences from proteins is statistically significant. Their analysis revealed that the value is located more than 3σ away from the mean RMSD calculated out of 10^6 superimpositions of same length randomly chosen peptides out of 32 proteins. Thus, the structural similarity of the modules, namely ten DNA recognition HTH modules and five DNA non-recognition HTH modules, is not likely to be a chance event. These DNA recognition and non-recognition modules are probably related in an evolutionary manner.

II-4. Sequence comparison among modules

Sequence comparison is performed on the basis of the 3D structural superimposition. The comparison is carried out only amongst structurally similar modules. The best fitted super-imposition of two modules is expressed on the sequence by aligning corresponding residues in their 3D structures. From that alignment, a similarity score is calculated. The score for each amino acid residue pair is obtained from the score matrix (Dayhoff *et al.*, 1978). Similarity of each set of modules is expressed by the summation of scores assigned to each pair of residues. Significance of the summation obtained for each pair of the sequences is examined against the value obtained between one sequence of the pair and randomly generated sequences. The random sequence takes into

Table II-1: Structural correspondence between helix-turn-helix modules found in prokaryotic repressors[a]

DNA-binding HTH modules			Other HTH modules			RMSD (Å)
2CRO	20-32	(M3)				0.00
1CRO	19-31	(~M3)				0.48
3FIS	77-89	(~M5)				0.57
1LCD	9-21	(~M2)				0.57
2TCT	30-42	(~M3)				0.60
2OR1	20-32	(~M2)				0.62
1BIA	25-37	(~M3)				0.63
			2CRO	48-60	(~M5)	0.64
			2OR1	48-60	(~M4)	0.65
2WRP	71-83	(~M5)				0.65
			1LRD	65-77	(~M5)	0.66
3GAP	172-184	(~M3)				0.67
1LRD	36-48	(~M3)				0.68
			2OR1	9-21	(~M1)	1.20
			2CRO	9-21	(~M2)	1.28

[a]Abbreviations used: 1BIA, *bio* repressor; 1CRO , λ Cro; 2CRO, 434 Cro protein; 3FIS, fis; 3GAP, catabolite gene activator; 1LCD, *lac* repressor; 1LRD, λ repressor; 2OR1, 434 repressor; 2TCT, Tet repressor; 2WRP, *trp* repressor.

Fig II-3. Superimposition of similar modules. By colour, red is 434 Cro protein module M5, yellow is 434 repressor module M2, cyan is 434 Cro protein module M5, light purple is 434 repressor module M4, slate is trp repressor module M5, light green is λ repressor module M5, deep blue is catabolite gene activator module M3, gray is λ repressor module M3, deep purple is 434 repressor module M1 and brown is 434 Cro protein module M2.

account the mean percentage of amino acid residues in proteins. A comparison with random sequence is carried out 10^4 times. With the mean value (m) and the standard deviation (σ) of a summation of scores obtained by 10^4 comparisons, significance (d) of the value for the pair of the sequence (s) is estimated as,

$$d = \frac{s - m}{\sigma} .$$

Sequence similarity of the 15 structurally similar modules was investigated. Among the ten DNA recognition HTH modules, the values lie approximately between 2σ and 4σ, except for a few cases. The values between a DNA non-recognition HTH module and a DNA recognition HTH module lie around 2.3σ (Table II-2) . The resemblance between a DNA non-recognition HTH module and a DNA base-recognition HTH module is sometimes at the same level as the resemblance of DNA base-recognition HTH modules. Since an evolutionary relationship was suggested by a sequence analysis of the motif (Ohlendorf *et al.*, 1983), some of the 15 modules, despite differences in functions may also be evolutionarily related. However, the possibility that the sequence similarities arose from the amino acid preference for α helices and turn formation cannot be ruled out.

II-5. Repeat of a helix-turn-helix module

Our results show that the ten investigated prokaryotic transcriptional regulators have one to three structurally similar HTH modules. Some function as a DNA base recognition interface and constitute the HTH motif. The DNA-binding proteins are made essentially of HTH modules, in repeat (Figure II-2).

Table II-2: Significance of similarity between the two sequences

	2CRO (M3)	2OR1 (~M2)	1LRD (~M3)	1CRO (~M3)	3GAP (~M3)	2WRP (~M5)	3FIS (~M5)	1LCD (~M2)	1BIA (~M3)	2TCT (~M3)	2CRO (~M5)	2OR1 (~M4)	1LRD (~M5)	2CRO (~M2)	2OR1 (~M1)
2CRO (M3)															
2OR1 (~M2)	5.30														
1LRD (~M3)	3.34	3.99													
1CRO (~M3)	2.70	3.05	3.38												
3GAP (~M3)	2.44	3.92	3.14	2.11											
2WRP (~M5)	2.11	3.05	3.15	2.64	2.27										
3FIS (~M5)	2.52		2.67	2.79	2.94										
1LCD(~M2)	2.77	2.67	2.53		3.32	2.07									
1BIA(~M3)	3.36	4.19	4.70	4.16	3.57	4.16	3.78	2.48							
2TCT(~M3)	3.81	3.79	3.74	3.26	2.43	3.20	3.26		4.22						
2CRO (~M5)					2.67		2.06			2.17					
2OR1 (~M4)									3.05	3.42					
1LRD (~M5)		2.21		2.11	2.36	2.93			2.86	2.39					
2CRO (~M2)			2.15	2.03			2.07		2.61	2.97			2.47		
2OR1 (~M1)			2.87	3.18		2.70	3.29		2.39					5.83	

The value is calculated out of a similarity score of two sequences in a column and a row subtracted by the mean score between the sequence in the row and random sequences and scaled by the standard deviation. The table is approximately symmetrical so that only half of the table is shown. Scores more than 2.06 which corresponds to an approximately 0.023 probability of chance appearance are shown.

Evolution of repressors by repetitive use of HTH modules

The similarity of the 15 modules suggests that a single HTH module might have multiplied to build up DNA-binding proteins. On the basis of organization of these proteins with HTH modules, one of the possible pathways of evolution of the transcription factors is suggested (Figure II-4). The pathway starts with a single HTH module that might bind DNA bases. One way leads to proteins without duplication of HTH modules; they are catabolite gene activator, *trp* repressor, λ Cro, fis, *lac* repressor, *bio* repressor and Tet repressor. The other way leads to a duplication after adding one connection module on the COOH-terminus. The pathway makes another branch at this stage, one of which leads to λ repressor. The other makes one more duplication and adds an HTH module on the NH_2-terminus. This might be the basis for 434 repressor and 434 Cro protein.

The easiest way to obtain a protein that regulates gene transcription might be to multiply a functionally important unit. Presumably a single DNA base-recognition HTH module that is unstable by itself was multiplied to gain conformational stability in an early stage of molecular evolution. The multiplication of a short gene encoding a protein segment is much more probable than random elongation of a gene during evolution (Ohno, 1984). The possibility of multiplication of the HTH modules is also supported by finding a high similarity amongst their sequences. The DNA base-recognition HTH modules were suggested to be derived from a common precursor (Ohlendorf *et al.*, 1983). Some of the sequences of the DNA non-recognition HTH modules are much more similar to the DNA base-recognition HTH modules.

Location of phosphate-binding modules

We present here a novel repetitive occurrence of HTH modules in

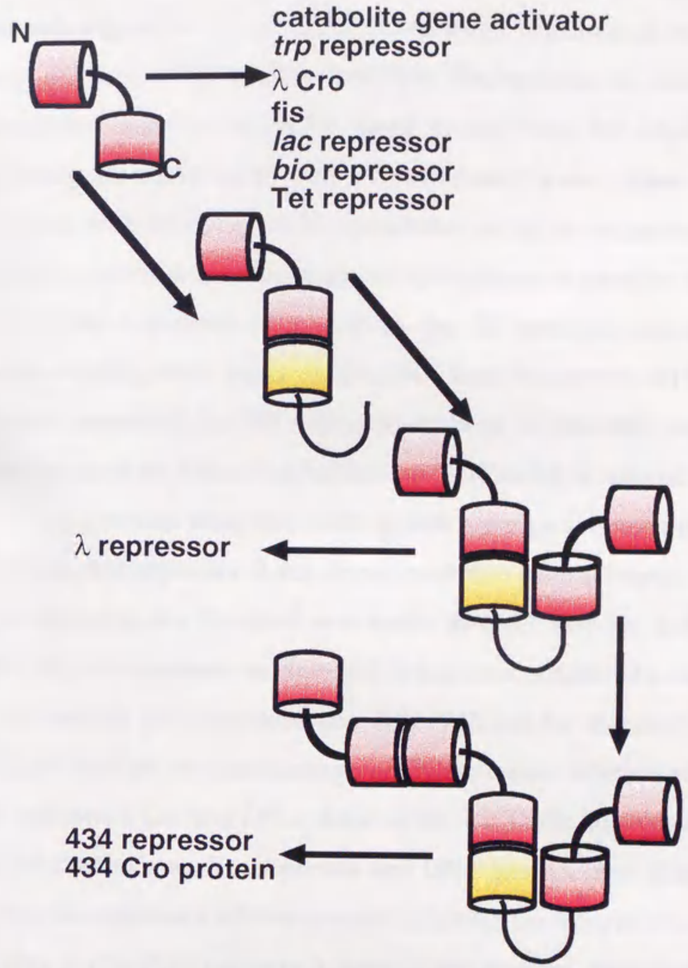

Fig II-4. Hypothetical reconstruction pathway of ten prokaryotic repressors with the HTH modules and others. A cylinder indicates a half-length helix. A red module is the HTH module. A yellow helix with a loop is a connector module. Repetitive utilization of the HTH modules and fusion of other modules relates to contemporary transcription factors. A conceivable pathway starts with a base-recognition HTH module. One branch leads to the present catabolite gene activator, *trp* repressor, λ Cro, fis, *lac* repressor, *bio* repressor and Tet repressor without duplication of HTH modules, while the other branch makes a duplication after adding one connective module on the COOH-terminal. The pathway branched into two at this stage, one of which leads to λ repressor. The other makes another duplication on the NH_2-terminal and becomes the basis of 434 repressor and 434 Cro protein.

prokaryotic repressors. The pattern of the module organization may facilitate a variety of these DNA-binding proteins. Recognition of DNA cannot be accomplished only by the HTH motif alone, since the recognition helix specifically interacts only with two or three bases. Other non-specific interactions with DNA backbone phosphates aid in the recognition. Although the DNA recognition sites are scattered throughout the proteins when they are shown on the sequences (Figure II-2), the 3D positions are similar. DNA phosphate-binding sites locate on the DNA base-recognition HTH module, on the module preceding the DNA base-recognition HTH module and around the NH_2-terminus of the helix lying behind the HTH motif, as viewed from a bound DNA. The consensus phosphate-binding sites amongst six proteins, namely 434 Cro protein, 434 repressor, λ repressor, catabolite gene activator, *trp* repressor and *lac* repressor are localized at a single module; they are 434 Cro protein module M4, 434 repressor module M3, λ repressor module M4, catabolite gene activator module M1, *trp* repressor module M3 and *lac* repressor module M3. We call this module the phosphate-recognition module. Interaction by hydrogen bonds between λ Cro and DNA, between fis and DNA, between *bio* repressor and DNA and between Tet repressor and DNA has not been determined. The phosphate-recognition modules always lie behind the recognition helix, except for λ Cro; λ Cro does not have a helix at the position. However, its relative position in space to the recognition helix and its position on a primary structure differ. For 434 Cro protein, 434 repressor , λ repressor that have two or three HTH modules and *lac* repressor, the phosphate-recognition module exists on the COOH-terminal side of the proteins and the structures are similar. For catabolite gene activator and *trp* repressor that have one HTH module, the phosphate-recognition module exists on the NH_2-terminal side of the proteins and here too the structures are similar. They resemble the structure of the HTH module, but

the RMSDs between the modules and the HTH modules are high. For fis, *bio* repressor and Tet repressor, the complex of the protein and a DNA has not been determined. There is an α helix behind the HTH motif attached to the NH_2-terminus, as in the case of catabolite gene activator and *trp* repressor. Although the structure of the module containing the α helix is not similar to the ones in catabolite gene activator and *trp* repressor, the topological appearance of the α helix is the same and, therefore, it might recognize the backbone of DNA.

The difference in the combination of modules gives variety to the position of the phosphate-recognition modules, which in turn makes differences in the manner of presenting the recognition helix to a major groove of DNA. Different organizations of the HTH modules may provide structural varieties required for sequence-specific DNA recognition. Classification of the ten proteins by the pattern of module combination, namely 434 Cro protein, 434 repressor, λ repressor and *lac* repressor in one group and catabolite gene activator, *trp* repressor, fis, *bio* repressor and Tet repressor in another is almost consistent with the hypothetical evolutionary pathway shown in Figure II-4. λ Cro is different from these two groups because it does not have an α helix behind the HTH module. It could be the result of early branching of λ Cro protein from catabolite gene activator, *trp* repressor fis, *bio* repressor and Tet repressor in the hypothetical reconstruction pathway in Figure II-4.

DNA-binding ability of HTH modules

A question may arise whether all the 15 similar HTH modules bind DNA, since ten of them do. Brennan and Matthews (1989) presented the mean amino acid change per codon (AAC) score to test whether a sequence is the HTH motif or not. They created a master set of HTH motifs and calculated a score of a tested segment against the set. The score is acquired by aligning the segment to

the master set without gaps, summing up the number of amino acid residues in the master set that differ from the segment in every aligned position and normalizing the value by dividing it by the number of amino acid residues in the master set. The lesser the score , the more likely the segment is to be HTH motif. They set the threshold at 0.80. Five HTH modules which do not recognize DNA bases do not pass the test. Although one could test the non-specific DNA-binding capability of these modules, experiments showed that the proteins bind non-specifically to DNA using the same HTH motif (Sauer *et al.*, 1990). There is a similarity in the alignment of the 15 modules and differences among the sequences (Figure II-5). Conservation of hydrophobic residues is apparent at positions -5, -1, 1 and 6 and a Gly at position 0. Phage 434 Cro protein module M5 and 434 repressor module M4 do not have residues with charges at positions 3 or 4, which are considered necessary for base recognition. λ repressor module M5 and 434 Cro protein module M2 have a Lys instead of Gly on the turn. The large side chain of Lys might hamper DNA binding. 434 Cro protein module M2 and 434 repressor module M1 have residues with a charge at positions -5 and -6 instead of hydrophobic residues. Because of these differences, the sequences do not pass the AAC value test.

One of the most obvious and common differences between the DNA base-recognition HTH modules and DNA non-recognition HTH modules in the alignment is that DNA base-recognition HTH modules have a residue with a small side chain at position 5 (red), whereas DNA non-recognition modules have one with a large side chain (blue). Site-directed mutagenesis analysis of the corresponding residue of the HTH motif of λ repressor revealed that an Ala substitution of the site does not change the ability of DNA binding (Hecht *et al.*, 1986) Single substitution experiments on Tet repressor showed that the smaller side chains at position 5 tended to retain the function (Baumeister *et al.*, 1992).

Fig II-5. Structural alignment of similar HTH modules shown in red in Fig II-2. An arrow at the bottom shows the range of superimposition. Similar amino acid residues are boxed. Outlined letters show residues at boundaries of the modules. The name on the left of the sequence indicates a protein and a module number. The module number corresponds to the sequence between outlined residues (the last outlined residue is included). A scale on the top is numbered as a Gly at the bend of the HTH motif becomes zero. Modules above the thick bar in the middle are DNA base-recognition HTH modules and below the bar are DNA non-recognition ones. The HTH motif starts at -9 and ends at 11 on 434 Cro protein module M3. Coloured residues are the noticeable difference between two kinds of module.

Table II-3. Energy of λ repressor helix-turn-helix module-DNA complex[a]

Type	Total[b]	Internal[c]	Binding[d]
Wild Gly46	27.02	30.63	-3.61
Gly46->Ala	33.97	37.65	-3.69
Gly46->Ser	33.59	35.82	-2.23
Gly46->Thr	42.59	45.59	-3.00
Gly46->Trp	87.52	88.19	-0.67
Gly46->Glu	74.56	74.68	-0.12

[a] The unit is kcal/mol.

[b] Total means whole energy of the module-DNA complex.

[c] Internal means internal energy of the module only.

[d] Binding was obtained from total minus internal.

Binding energy of λ repressor module M3 which composes the motif is calculated by BIOGRAF (Mayo *et al.*, 1990), as shown in Table II-3. The total energy of the wild type module is reduced by 3.61 kcal/mol on binding to DNA. When Gly at position 5 of λ repressor module M3 is replaced with a Trp, the energy is augmented to -0.67 kcal/mol with little gain in stability by DNA binding. When the residue is replaced with an Ala, the energy is -3.69 kcal/mol, almost the same as the wild type. A Ser mutation results in -2.23 kcal/mol and a Glu substitution ends in -0.12 kcal/mol. Therefore, a large side chain at position 5 seems to hamper DNA binding (Figure II-6).

This calculation apparently does not hold for the HTH motifs for most of the homeodomains of eukaryotes. They have residues with a large side chain at position 5 and bind DNA. However, the DNA complex of engrailed homeodomain (Kissinger *et al.*, 1990) revealed that DNA shifts approximately two turns of the recognition helix toward the COOH-terminus in the complex. As Kissinger *et al.* (1990) pointed out, the HTH motif of the homeodomain differs from the ones of prokaryote repressors and the difference is shown in Figure II-7. Residues interacting with DNA bases by hydrogen bonds are in red. The HTH motif of the homeodomain does not interact with DNA. DNA bases are recognized at the middle of the latter helix which is much longer than the corresponding one in the HTH motif of prokaryotic repressors. DNA probably interacts with the next module. The interaction seems to avoid contact between DNA and the large side chain at position 5 which is coloured in orange. Therefore, our analysis is consistent with the experimental observations that the HTH module with a large side chain at position 5 does not interact with DNA. The AAC value test detects the overall amino acid sequence pattern for DNA base recognition and helix formation. The size of the side chain at position 5 seems a good signature of the docking pattern with a DNA.

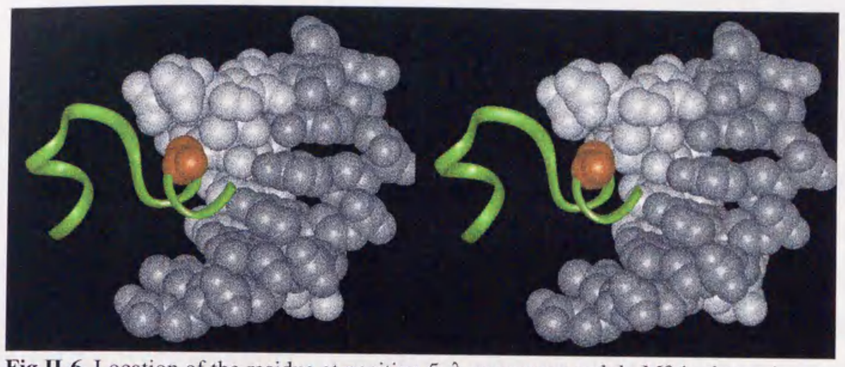

Fig II-6. Location of the residue at position 5. λ repressor module M3 is shown in green tube model with the residue at postion 5 (Gly) in orange space filling model. DNA is shown in gray and white. Gly is located close to DNA bases, so that substitution to a residue with a bulky side chain is not allowed.

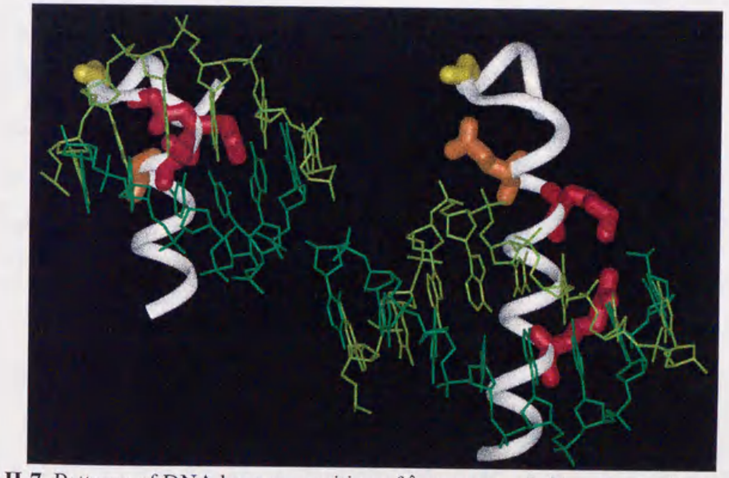

Fig II-7. Patterns of DNA base recognition of λ repressor (left) and engrailed homeodomain (right). Main chain traces are depicted by white tube models. Red residues with side chain form hydrogen bonds to DNA bases, yellow residues are glycines located at the turns and orange residues are those located at the five-residue COOH-terminal side of the glycine on the turn. Backbones of DNAs are drawn in green with thin wire model. DNA interaction sites of engrailed homeodomain were determined according to Kissinger et al (1990).

Physicochemical implication of the modules

Some of the physicochemical characters of fused modules were observed in experiments to test the stability of the protein fragment (Kelley & Yanofsky, 1985). They terminated the sequence of *trp* repressor at the 68th residue and found that the sequence from the first to the 68th residue interacts with other wild type *trp* repressor to form an inactive dimer. It is presumed that the sequence was capable of folding (Schevitz *et al.*, 1985). The 68th residue almost corresponds to the boundary of the fourth and fifth modules. In another experiment, *trp* repressor was cleaved at the COOH-terminal side of the 71st residue by chymotrypsin and the fragment from the eighth residue to the 71st residue had the native-like features, in the NMR spectrum (Tasayco & Carey, 1992). The cleavage site corresponds precisely to the boundary of the fourth and fifth modules. These experiments imply that the fused modules, namely fragments composed of module M1 through to module M4 could be a folding unit. Another NMR experiment on denatured 434 repressor noted a hydrophobic cluster between the 54th and 59th residues (Neri *et al.*, 1992). They discussed that the cluster might be an initial core for folding. This portion corresponds to the centre of module M4 and suggests that the single module could also be a folding core.

Chapter III.
Repetitive use of a phosphate-binding helix-turn-helix module in transcription factors and a repair enzyme

Abstract: Motifs for sequence specific protein-DNA interaction, such as helix-turn-helix, zinc finger, leucine zipper are now better understood as a result of extensive studies of three-dimensional structures of transcription factors. On the other hand, little attention has been paid to motifs for sequence non-specific binding, namely DNA phosphate binding. To address the question whether there is a case that different transcription factors and DNA manipulation enzymes, namely enzymes that work on DNA, share a similar mode of phosphate binding, we surveyed interactions between DNA and protein module, a structural unit of a globular protein. We analyzed the module organization of DNA polymerase β and found that residues making contact with DNA phosphates were localized to five modules. Structural comparison of these phosphate-binding modules against others in transcription factors and DNA manipulation enzymes revealed that DNA polymerase β, Oct-1 POU domain, 434 Cro protein and Arc repressor have a phosphate-binding module with three-dimensional structures similar to one another. This newly detected module, phosphate-binding helix-turn-helix (pbHTH) module named after its function and three-dimensional structure, interacts with DNA by making hydrogen bonds between a DNA phosphodiester oxygen and an amino hydrogen of the main chain located at the N-terminus of a C-terminal α helix, and making electrostatic interactions between DNA phosphates and side chains of lysine or arginine. Finding structurally and functionally similar phosphate-binding units in different transcription factors and DNA manipulation enzymes suggests that shuffling of modules is not limited to the DNA base-recognition module. Phosphate-binding modules are apparently also shuffled in DNA-binding proteins.

III-1. Protein-DNA non-specific interactions

Protein-DNA interactions are classified into sequence specific and non-specific interactions, the former being mainly achieved by interactions of protein side chains with DNA bases (Choo & Klug, 1997). Four kinds of bases in DNA have unique atom positions that can form hydrogen bonds to or electrostatic interactions with side chain atoms of amino acid residues. Non-specific interactions with DNA occur by interactions of main or side chains of proteins and DNA phosphates. Sugars of the DNA backbone have the potential to form hydrogen bonds to electron donors. When a main or a side chain of proteins forms a hydrogen bond to a sugar, the bond often bifurcates to a DNA phosphate (Mandel-Gutfreund *et al.*, 1995). As DNA has phosphodiester oxygens in its backbone, the existence of which does not depend on DNA sequences, proteins interact non-specifically with DNA phosphates. This interaction often aids in properly locating specific-binding sites of the protein onto DNA (Pabo & Sauer, 1992).

Protein-DNA specific interactions have been thoroughly examined and common three-dimensional (3D) structures in interaction regions have been identified. The well known DNA-binding motifs are helix-turn-helix (HTH), zinc finger, basic leucine zipper and basic helix-loop-helix motifs. These motifs which have one α helix inserted into a major groove of DNA "read" DNA sequences (Harrison, 1991). Sequence specific interaction on a minor groove found in purine repressor (Schumacher *et al.*, 1994) utilizes the α helix, and in TATA-box binding protein (Tan & Richmond, 1998) and integration host factor (Rice, 1997) both utilize β-sheet.

Unlike DNA sequence specific-binding motifs, a structural similarity in DNA non-specific binding regions has not been found. Pabo and Sauer (1992) described the situation as; "there does not appear to be any simple "rule" or pattern describing which residues are used for backbone contact."

We reported that the HTH motif is composed of a single module (Yura *et*

al., 1993). A globular domain was found to be partitioned into compact sub-structures, modules (Gō, 1981). Average length of the module is about 15 residues. Correspondence of module boundaries and intron positions of the genes of hemoglobin, lysozyme and other proteins was reported (Gō, 1981; Gō, 1983; Gō, 1985; Gō & Nosaka, 1987; Gilbert & Glynias, 1993; Tittiger *et al.*, 1993; Gō & Noguti, 1995; de Souza *et al.*, 1996). Thus, module is likely to be a unit of protein evolution. Module is also a unit of function, as it was shown experimentally by RNase activity in single modules (Yanagawa *et al.*, 1993), by exchanging modules of hemoglobin α and β subunits resulting in a swap of the function (Wakasugi *et al.*, 1994; Inaba *et al.*, 1998), and by exchanging a single module in isocitrate dehydrogenase that binds NADP to a module in NAD-specific isopropylmalate dehydrogenase resulting in exchange of coenzyme specificity without loss of activity (Yaoi *et al.*, 1996). In HTH proteins, common HTH modules recognize specific DNA sequences (Yura *et al.*, 1993).

Is there a shared structural unit for DNA non-specific binding, namely DNA-phosphate binding? We made use of module structure of transcription factors and DNA manipulation enzymes including modification, restriction, repair, replication enzymes and polymerases to address the question. It is ideal to find 3D structures of proteins that bind DNA, without sequence specificity. DNA polymerase β is one of the best targets to analyze how a protein non-specifically binds to DNA. DNA polymerase β was found to interact with DNA mostly by sequence non-specific hydrogen bonds (Pelletier *et al.*, 1994). As phosphate-binding modules were found in DNA polymerase β, we searched for similar modules in other transcription factors and DNA manipulation enzymes.

III-2. Module organization and protein function

Module boundary determination

Module boundaries were determined by centripetal and extension profiles (Gō & Nosaka, 1987; Noguti *et al.*, 1993). Candidates of module boundaries were obtained by locating local minima of F_i. F_i is an index of centripetal character of the ith residue, calculated by a mean-square distance between Cα atom of the ith residue and Cα atoms within a window of $(2k+1)$ residues (k = 15, 20, 25, 30, 35, 40 and 45) centered by the ith residue. The centripetal profile detects the local center of a protein. The candidates were confirmed by extension profile that checked the extendedness of the chain around the candidate module boundaries. N- and C-terminal module boundaries were determined separately, because one could not set large k at the terminus. Final determination of module boundaries was automatized (Sato *et al.*, in preparation). The treatment of the N- and C-termini of a protein resulted in detection of a new module boundary in 434 Cro protein. It was reported to have five modules by Yura *et al.* (1993), but the C-terminal module decomposed into two (Figure III-1).

Defining a protein-DNA contact

Heavy atoms in protein and DNA were defined to be in contact, when located within the range of 4.0Å distance. An amino acid residue and a nucleotide were defined to be in contact when three or more pairs of contact were found between them. These cutoffs were set to choose residues that were obviously in contact with DNA. DNA were divided into two parts; base and phosphate plus sugar. When atoms in a residue were in contact with a base, the residue was defined to be in contact with DNA base, even when other atoms in the same residue were in contact with phosphate and/or sugar. When atoms in a residue were in contact with phosphate and/or sugar but not with base, the

residue was defined to be in contact with DNA phosphate. A module that contained a base-contact residue was called base-recognition module, and that containing only a phosphate-contact residue was called phosphate-binding module.

Analysis of the module-DNA hydrogen bond

Since locations of hydrogens are usually not given in the Protein Data Bank (Bernstein *et al.*, 1977), they were geometrically calculated. Hydrogen bond analysis was based on the method of Baker and Hubbard (1984).

DNA phosphate-binding and ion-binding modules in DNA polymerase β

A black arrowhead in Figure III-1A indicates a module boundary of DNA polymerase β. There are 24 modules in DNA polymerase β. Length of a module in DNA polymerase β is about 14 residues on average. Module organization in 3D structure of DNA polymerase β is depicted in Figure. III-2. DNA polymerase β has been subdivided into four subdomains, based on structural comparisons among nucleic acid polymerases (Pelletier *et al.*, 1994). The N-terminal 8-kDa subdomain is composed of five modules, the fingers subdomain of five modules, the palm subdomain of seven, and the thumb subdomain of seven (Figure III-3). The subdomain junctions approximately correspond to module boundaries.

DNA polymerase β made contacts with DNA, as shown in Table III-1. The contact was localized to seven modules, M2, M4, M5, M7, M16, M19 and M20. Modules M4, M5, M7, M16 and M20 make contact with DNA phosphates and sugars, but not with DNA bases. These five modules can be called phosphate-binding modules. Modules M2 and M19 made contact with DNA phosphates and sugars and also with DNA bases. When an amino acid residue contacts a DNA base located deep in a groove, the same residue inevitably contacts DNA phosphates and sugars located at sides of the groove. Therefore,

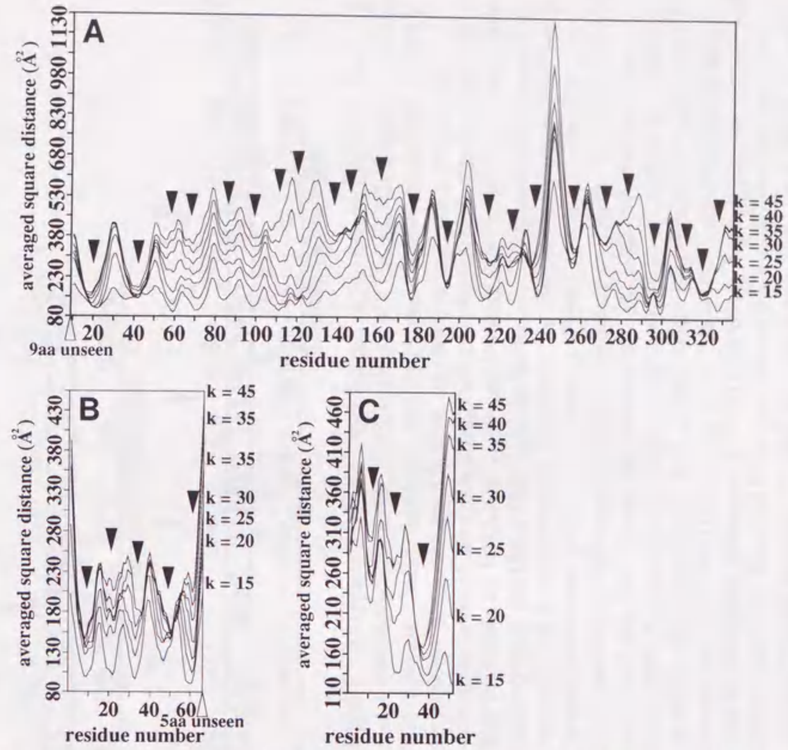

Fig III-1. Centripetal profiles of (A) human DNA polymerase β (PDB code: 1BPY) (Sawaya et al., 1997), (B) 434 phage 434 Cro protein (3CRO) (Mondragon & Harrison, 1991) and (C) P22 phage Arc repressor (1PAR) (Raumann et al., 1994). Module boundaries are described by balck arrowheads. Each profile has seven lines that corresponded to seven values of k, a window length of Cα atoms around ith Cα atom. Nine residues at the N-terminus of of DNA polymerase β and five residues at the C-terminus of 434 Cro protein were not observed by X-ray crystallographic analysis (Sawaya et al., 1997; Mondragon & Harrison, 1991).

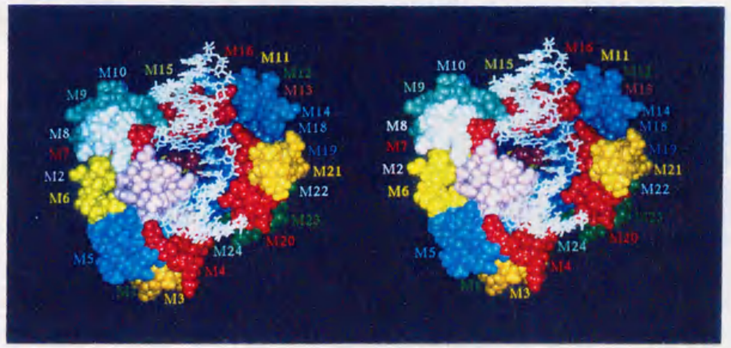

Fig III-2. Module organization of DNA polymerase β (PDB code:1BPY) (Sawaya et al., 1997). DNA is depicted by stick model in white. The active site residues are on brown module M13 and dark blue module M17 at the center. Phosphodiester oxygens are mainly bound to modules in red, and in purple at the left bottom corner.

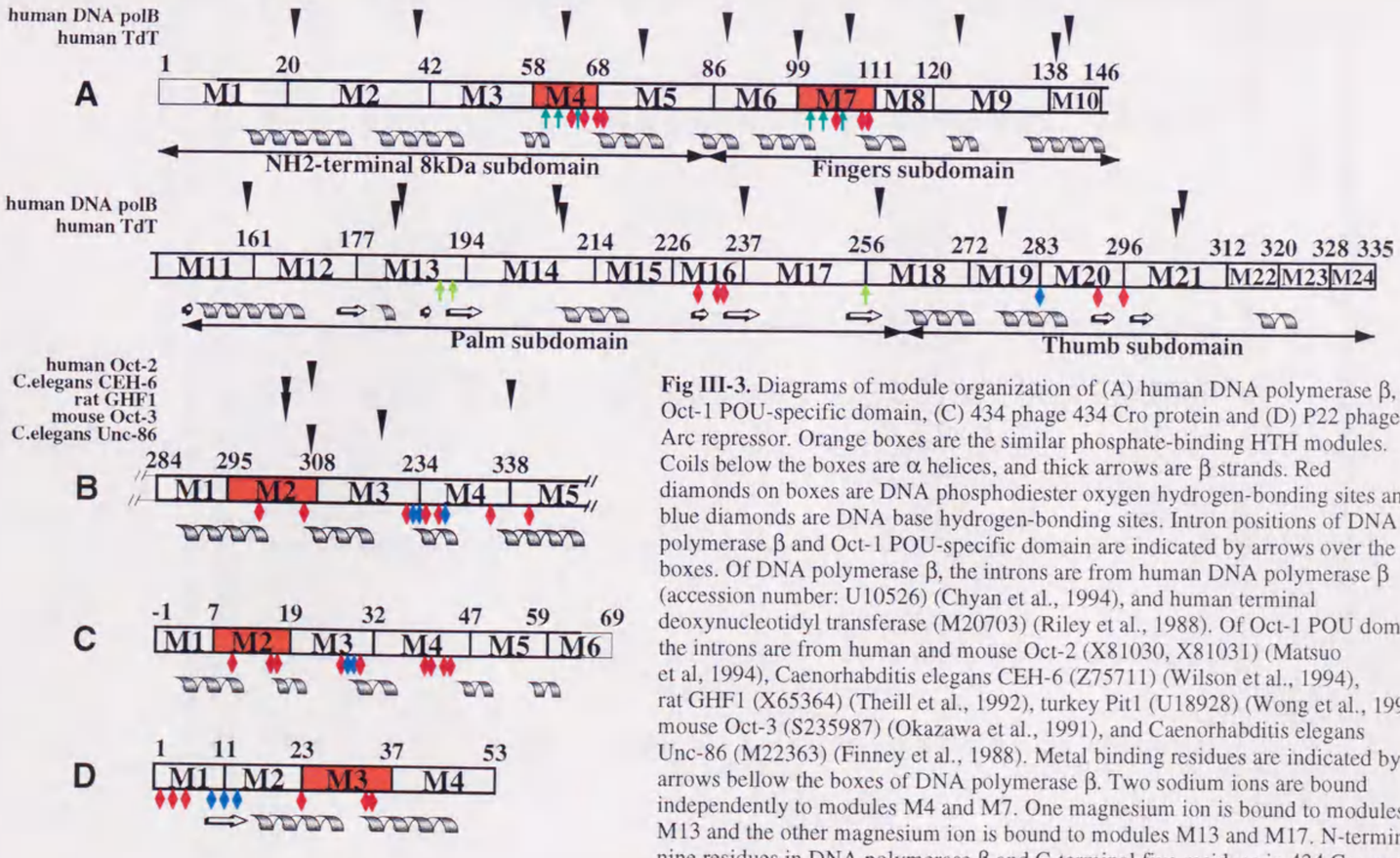

Fig III-3. Diagrams of module organization of (A) human DNA polymerase β, (B) Oct-1 POU-specific domain, (C) 434 phage 434 Cro protein and (D) P22 phage Arc repressor. Orange boxes are the similar phosphate-binding HTH modules. Coils below the boxes are α helices, and thick arrows are β strands. Red diamonds on boxes are DNA phosphodiester oxygen hydrogen-bonding sites and blue diamonds are DNA base hydrogen-bonding sites. Intron positions of DNA polymerase β and Oct-1 POU-specific domain are indicated by arrows over the boxes. Of DNA polymerase β, the introns are from human DNA polymerase β (accession number: U10526) (Chyan et al., 1994), and human terminal deoxynucleotidyl transferase (M20703) (Riley et al., 1988). Of Oct-1 POU domain, the introns are from human and mouse Oct-2 (X81030, X81031) (Matsuo et al, 1994), Caenorhabditis elegans CEH-6 (Z75711) (Wilson et al., 1994), rat GHF1 (X65364) (Theill et al., 1992), turkey Pit1 (U18928) (Wong et al., 1992), mouse Oct-3 (S235987) (Okazawa et al., 1991), and Caenorhabditis elegans Unc-86 (M22363) (Finney et al., 1988). Metal binding residues are indicated by arrows bellow the boxes of DNA polymerase β. Two sodium ions are bound independently to modules M4 and M7. One magnesium ion is bound to modules M13 and the other magnesium ion is bound to modules M13 and M17. N-terminal nine residues in DNA polymerase β and C-terminal five residues in 434 Cro protein shown in thin lines were not determined by X-ray crystallography (Sawaya et al., 1997; Mondragon & Harrison, 1991). For Oct-1,the POU-specific domain only is shown.

Table III-1: Residues of DNA polymerase β in contact with DNA

DNA polymerase β		DNA		no. of contacts		
module	residue	chain	nt	phosphate	sugar	base
M2	His34	T	C5	0	0	23
M2	His34	D	G1	0	0	3
M2	Lys35	D	G1	5	2	0
M2	Ala38	D	G1	0	2	1
M2	Tyr39	D	G1	1	3	0
M4	Gly64	D	T2	0	7	0
M4	Gly64	D	C3	4	1	0
M4	Val65	D	C3	4	0	0
M4	Gly66	D	T2	8	4	0
M4	Thr67	D	T2	3	0	0
M4	Lys68	D	G1	11	2	0
M4	Lys68	D	T2	7	0	0
M5	Ile69	D	T2	7	0	0
M7	Gly105	P	C8	0	5	0
M7	Gly105	P	G9	3	0	0
M7	Ile106	P	G9	3	0	0
M7	Gly107	P	C8	10	3	0
M7	Pro108	P	C8	5	0	0
M7	Ser109	P	G7	0	4	0
M7	Ser109	P	C8	8	0	0
M7	Ala110	P	C8	3	0	0
M16	Ser229	T	A11	4	1	0
M16	Lys230	T	A11	4	3	0
M16	Gly231	T	C10	5	5	0
M16	Thr233	T	C10	7	2	0
M16	Lys234	T	C10	4	2	0
M19	Lys280	T	G6	7	1	5
M19	Arg283	T	G6	0	5	5
M19	Arg283	T	G7	0	4	1
M20	Leu287	T	G7	3	0	0
M20	Thr292	T	G7	2	1	0
M20	Glu295	T	C8	1	10	0

Two atoms within 4.0Å were considered to be in contact. A residue with less than three pairs of contacts was omitted. Chains in the DNA are depicted as D (downstream), T (template) and P (primer).

modules M2 and M19 were not regarded as phosphate-binding modules. Modules M4, M5, M7, M16 and M20 were placed around DNA, as if to grasp DNA (Figure III-2 in red, except for M5 in purple).

In DNA polymerase β, two sodium ions and two magnesium ions bind to become an active enzyme. Based on crystal soaking experiments, sodium ions could be potassium ions (Pelletier *et al.*, 1996). Residues that binds metal ions are depicted in Figure III-3. One of the sodium ions was exclusively bound to module M4, while the other was bound to module M7. Modules M4 and M7 are Na^{+}-binding modules. These two modules are at the same time phosphate-binding modules and their 3D structures are strikingly similar (Figure III-4). One of the magnesium ions was bound to oxygen atoms of two aspartate residues in module M13. The other magnesium ion was bound to the same aspartate residues with different oxygen atoms of the side chains and to an aspartate residue on module M17. Modules M13 and M17 are Mg^{2+}-binding modules. Even though these two modules had different 3D structures, they use oxygen atoms of aspartate residues located close to the C-terminal of the modules to coordinate Mg^{2+}.

III-3. Structurally similar phosphate-binding modules

Superimposition of modules

Three-dimensional structure of modules that interact with DNA phosphate in all the DNA-binding proteins in Protein Data Bank (Bernstein *et al.*, 1977) were compared by superimposing the Cα atoms. If length of amino acid sequences differed between two modules, all possible sequential correspondence of Cα atoms of the two modules was considered.

In the present work, modules with helices on both termini were treated. Cutoff for the similarity was determined to be around 2.0Å. In 574 proteins of

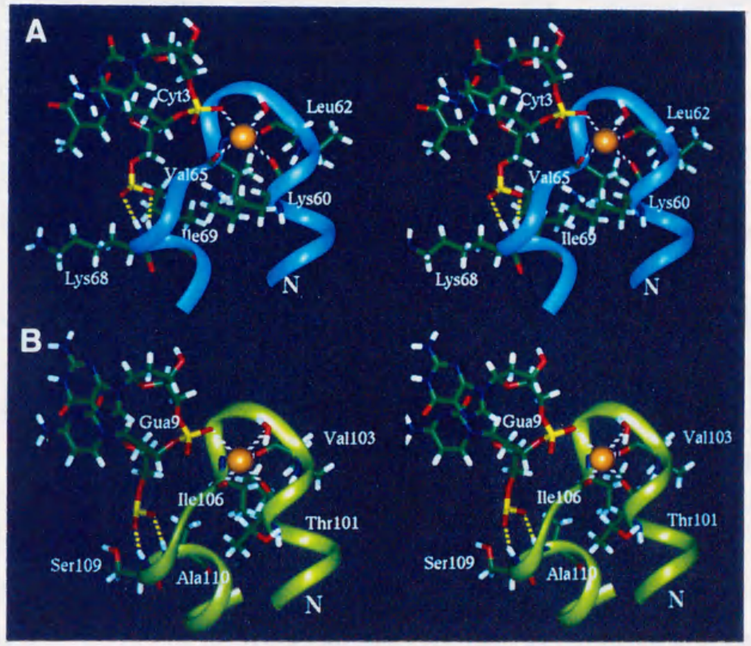

Fig III-4. Na^+-binding modules in DNA polymerase β. (A) module M4 and (B) module M7. Orange atoms are sodium. Residues that have atoms to coordinate the sodium atom are shown in the stick model. Pink dotted lines indicate coordinations. Phosphodiester oxygens of DNA make hydrogen bonds to α helices at the C-terminus, as shown in yellow dotted lines. Residues with donor and acceptor atoms are also shown in the stick model.

sequence identity less than 30% in Protein Data Bank (Bernstein *et al.*, 1977), root mean square deviation (RMSD) calculation was resulted in 5.04Å on average with 1.62Å of standard deviation, when 1,387 modules with a helices on both termini were randomly chosen and superimposed in every possible pair (data not shown). RMSD less than about 2.0Å falls within 3% of the entire number of module pairs. Therefore, two modules with RMSD less than 2.0Å are similar, with a statistical significance at the 3% level. Wintjens *et al.*(1996) evaluated the significance of RMSD of a protein-fragment with two α helices connected by a β turn structure and found that 2.5Å was the reasonable cutoff for classification of peptides of 20 amino acid residues.

A search for a similar phosphate-binding module in DNA-binding proteins

Thirty-nine transcription factors and DNA manipulation enzymes of which 3D structures were solved with DNA were decomposed into modules. Modules that contact DNA phosphates were compared against the five phosphate-binding modules of DNA polymerase β. As a result, module M2 of Oct-1 POU domain (Klemm *et al.*, 1994), module M2 of 434 Cro protein (Mondragon & Harrison, 1991) and module M3 of Arc repressor (Raumann *et al.*, 1994b) were found to have structures similar to modules M4 and M7 of DNA polymerase β (Figures III-5 and III-6A). The similarity in 3D structures of those modules was not apparent out of their sequence comparison (Figure III-6B). DNA polymerase β modules M4 and M7 coincided with helix-hairpin-helix (HhH) sequence motif, one of the DNA-binding motifs first found in endonuclease III (Pelletier *et al.*, 1996; Thayer *et al.*, 1995; Seeberg *et al.*, 1995; Doherty *et al.*, 1996). Amino acid sequence search, however, could not locate structurally/functionally similar unit in Oct-1 POU domain, 434 Cro protein and Arc repressor. The pbHTH module turned out to include HhH motif and also to include modules with a structural/functional similarity without sequence similarity to HhH motif.

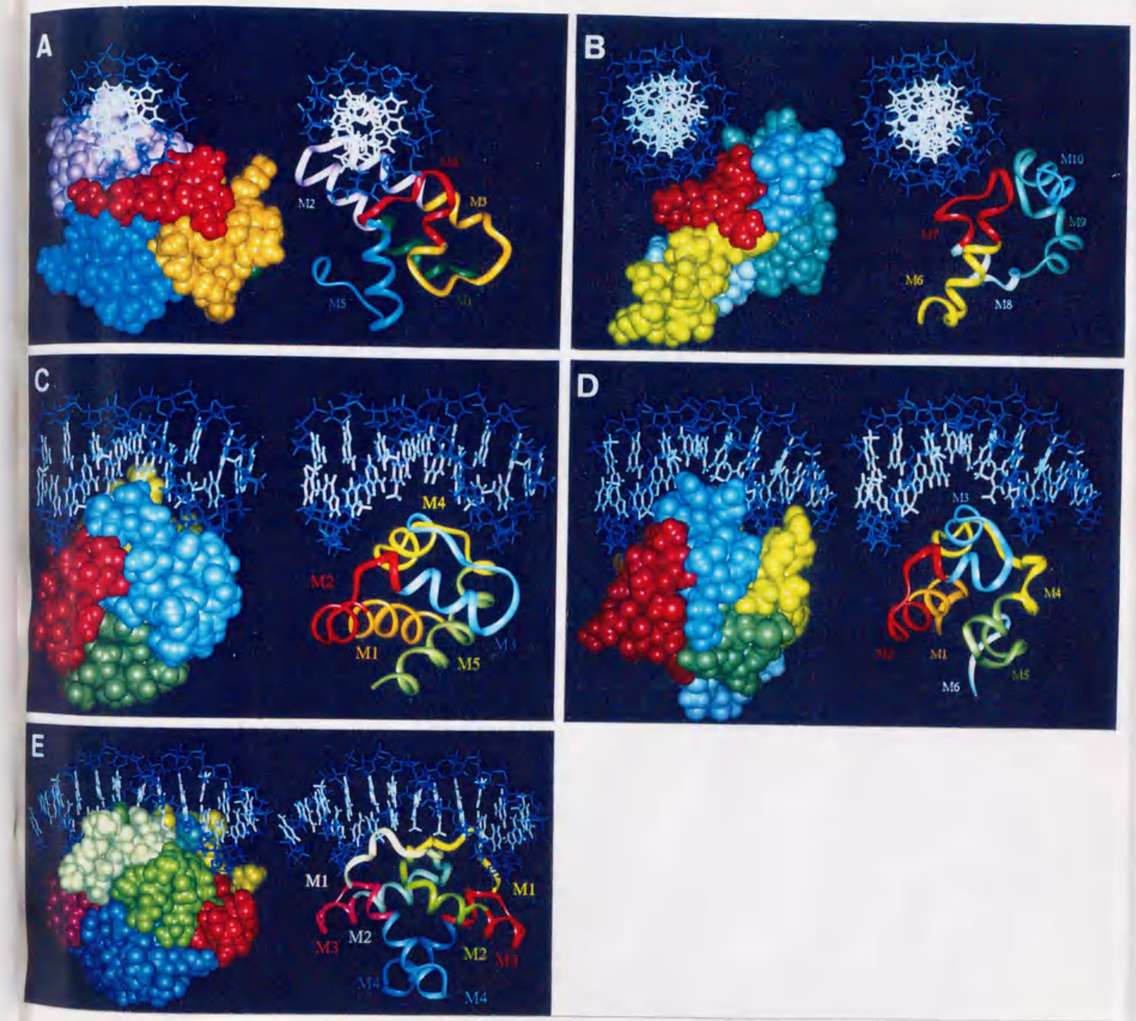

Fig III-5. Module organization of (A) DNA polymerase β from modules M1 to M5, (B) DNA polymerase β from modules M6 to M10, (C) Oct-1 POU-specific domain, (D) 434 Cro protein monomer and (E) Arc repressor dimer depicted by space filling (left) and tube (right) models. Each protein is colored by a module. DNA are depicted by the stick model. DNA backbones are colored blue and bases are in white.

B

```
DNA pol β (M4)  56 GAEAKKLPGVGTKIAEK  73
DNA pol β (M7)  97 INFLTRVSGIGPSAARK 113
Oct-1     (M2) 295 FKQRRIKLGFTQGDVGL 311
434Cro    (M2)   6 LKKRRIALKMTQTELAT  22
Arc rep   (M3)  22 VRKVAEENGRSVNSEIY  37
                   helix     helix
                   Superimposed 13 residues
```

Fig III-6. (A) Superimposition of phosphate-binding HTH modules. Each module is depicted by the Cα trace tube model. White regions are part of the previous and next modules. Light blue indicates the N-terminal side. DNA polymerase β modules M4 and M7 are in blue and green, respectively, Oct-1 module M2 in purple, 434 Cro module M2 in red and Arc repressor module M3 in yellow. (B) Structural alignment of phosphate-binding HTH modules in (A). Amino acid residues in bold fonts are included in the module. Residues in thin fonts are ones of juxtaposing modules. Blue residue has positive charge on a side chain and green residue has a small side chain located at one residue before the C-terminal a helix.

Oct-1 and 434 Cro protein are HTH proteins. The overall structural similarity of the Oct-1 POU-specific domain and 434 Cro protein became known when the 3D structure of Oct-1 was determined (Assa-Munt *et al.*, 1993). The Oct-1 POU-specific domain was decomposed into five modules (Figures III-3 and III-5C). Details of the module organization of Oct-1 will be discussed elsewhere (Yura *et al.*, in preparation).

434 Cro protein was decomposed into six modules (Figure III-1B). Modules M2 and M4 are phosphate-binding modules while module M3 is a base-recognition module (Figure III-3). The spatial arrangement of modules is depicted in Figure III-5D. 434 Cro protein and 434 repressor (Aggarwal *et al.*, 1988) essentially have the similar 3D structures. 434 repressor has the phosphate-binding modules and base-recognition module corresponding to those of 434 Cro (Yura *et al.*, 1993).

The Arc repressor has a DNA-binding ribbon-helix-helix (RHH) motif (Raumann *et al.*, 1994a). DNA bases are recognized by an anti-parallel β-sheet created by dimerization. The monomer decomposed into four modules (Figures III-1C and III-5E) and DNA phosphates were mainly bound to modules M1 and M3 (Figure III-3). Module M1 was hydrogen bonded to bases and phosphodiester oxygens, whereas module M3 was hydrogen bonded only to phosphodiester oxygens.

We reported that 434 Cro protein had three modules, M2, M3 and M5, that were structurally similar (Table III-2). Module M3 was almost equivalent to the HTH motif (Yura *et al.*, 1993). Hence, we named modules M4 and M7 of DNA polymerase β, M2 of Oct-1 POU domain, M2 of 434 Cro protein and M3 of Arc repressor, phosphate-binding HTH (pbHTH) module.

Table III-2: Root mean square deviation of HTH modules (Å)

module	M4 (pol)	M7 (pol)	M2 (Oct)	M2 (434)	M3 (Arc)	M3 (434)	function
M4 (DNA pol. β)							phosphate
M7 (DNA pol. β)	0.60						phosphate
M2 (Oct 1)	2.46	2.46					phosphate
M2 (434 Cro)	2.38	2.32	0.66				phosphate
M3 (Arc rep.)	2.18	2.04	1.29	0.99			phosphate
M3 (434 Cro)	2.08	2.06	1.53	1.30	0.61		base
M5 (434 Cro)	2.38	2.30	1.51	1.26	0.71	0.71	scaffold

Root mean square deviation was calculated by taking the average Cα deviation of thirteen residues. Correspondence of residues is shown in Figure III-6B. Phosphate in the column of function indicates that the module is a pbHTH module, base indicates a brHTH module and scaffold indicates an sfHTH module.

III-4. Phosphate-binding HTH module

Characteristics of pbHTH module

The five modules, M4 and M7 of DNA polymerase β, M2 of Oct-1 POU domain, M2 of 434 Cro protein and M3 of Arc repressor, bound a DNA phosphate, in a similar manner and all used backbone amino hydrogens (NHs) at the N-terminus of the C-terminal α helix of the module. An α helix is formed by hydrogen bonds between a backbone carbonyl oxygen (CO) belonging to *i*th residue and an NH of *i*+3rd residue. At the N-terminus of the α helix, however, three NHs are free from making hydrogen bonds with CO atoms. The free NH has the potential to form hydrogen bonds with a hydrogen acceptor. DNA phosphodiester oxygen formed a hydrogen bond to one of those NHs in these five modules.

In modules M4 and M5 of DNA polymerase β, there were three free NHs, namely Thr67, Lys68 and Ile69 at the N-terminus of the C-terminal α helix of

Table III-3: Hydrogen bonds between phosphate-binding HTH modules and DNA

	Protein			DNA		
	module	residue	atom	chain	nt	atom
DNA polymerase β	M4	Gly64	**N**	D	C3	**O2P**
	M4	Gly66	**N**	D	T2	**O2P**
	M4	Lys68	**N**	D	T2	**O1P**
	M4	Lys68	Nζ	D	G1	**O1P**
	M5	Ile69	**N**	D	T2	**O2P**
	M7	Gly105	**N**	P	G9	**O1P**
	M7	Ser109	**N**	P	C8	**O2P**
	M7	Ala110	**N**	P	C8	**O1P**
Oct-1	M2	Arg299	Nη2	A	T3	**O1P**
	M2	Gln306	**N**	A	T3	**O2P**
434 Cro	M2	Arg10	Nη1	A	T4	**O2P**
	M2	Thr16	Oγ1	B	T3	**O1P**
	M2	Gln17	**N**	B	T4	**O1P**
Arc repressor	M3	Val33	**N**	E	A4	**O1P**
	M3	Asn34	**N**	E	A4	**O2P**

The atoms in the left column form hydrogen bonds to atoms in the right column in the same row. Atoms of protein and DNA backbones are written in bold fonts to emphasize that most bonds are between backbones. Chain identifications of the DNAs are as the ones in the Protein Data Bank (Bernstein *et al.*, 1977).

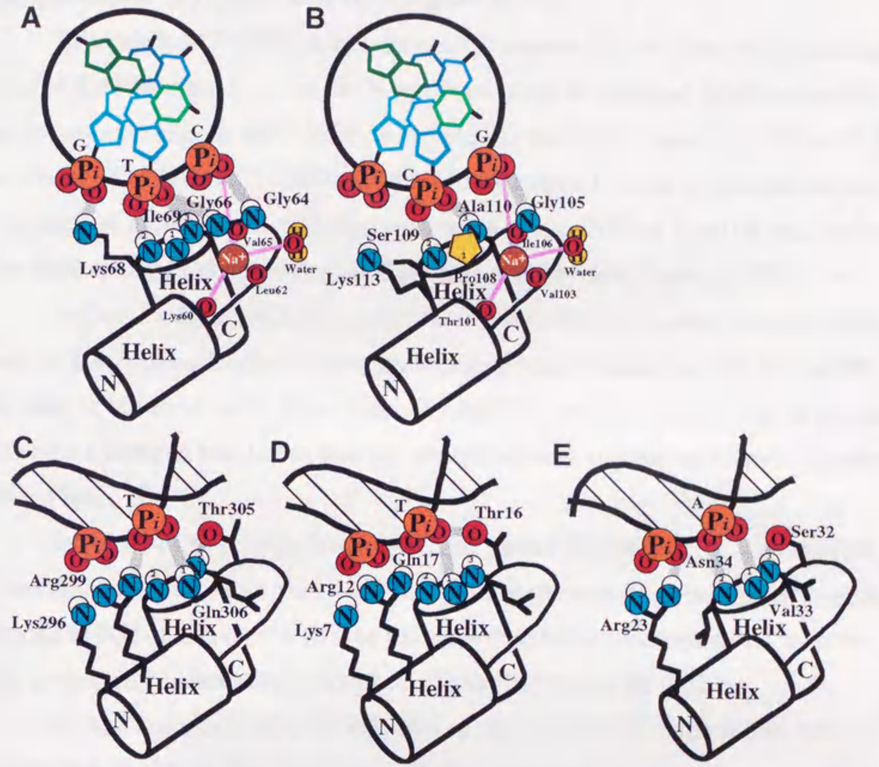

Fig III-7. A Schematic representation of DNA-pbHTH module interactions. (A) DNA polymerase β module M4, (B) DNA polymerase β module M7, (C) Oct-1 POU domain module M2, (D) 434 Cro protein module M2 and (E) Arc repressor module M3. One or two of the backbone NHs at the N-terminal of the C-terminal α helix form hydrogen bonds to phosphodiester oxygens of DNA (gray lines). Lysines or arginines are located on the N-terminal helix in case of Oct-1 POU domain M2, 434 Cro protein M2 and Arc repressor M3. They seem to interact electrostatically with DNA phosphodiester oxygens. Lysines were found on the C-terminal helix in case of DNA polymerase β M4 and M7. In (A) and (B), sodium ions are shown in orange. Coordinations are indicated by pink dotted lines.

module M4. NHs of Lys68 and Ile69 formed hydrogen bonds to phosphodiester oxygens of thymine. In addition to these hydrogen bonds, main chain NHs of Gly64 and Gly66 and the side chain of Lys68 hydrogen bonded to DNA phosphodiester oxygens (Table III-3; Figure III-7A).

In module M7 of DNA polymerase β, there were two free NHs, namely NHs of Ser109 and Ala110 at the N-terminus of the C-terminal α helix. Residue at 108 was proline; a side chain of proline is covalently bonded to NH of the backbone. The NHs of Ser109 and Ala110 hydrogen bonded to phosphodiester oxygens of cytosine. In addition, the main chain NH of Gly105 hydrogen bonded to phosphodiester oxygen of guanine (Table III-3; Figure III-7B).

In Oct-1 module M2, two out of three NHs of the C-terminal α helix were free. A DNA phosphodiester oxygen was hydrogen bonded to NH of Gln306, the first of the three NHs. Side chain of Arg299, a residue on the other α helix, was also hydrogen bonded to another phosphodiester oxygen of thymine (Table III-3; Figure III-7C).

In the 434 Cro protein module M2, all three NHs on the N-terminus of the C-terminal α helix were free. A DNA phosphodiester oxygen was hydrogen bonded to NH of Gln17, the first of the three free NHs. Side chain of Thr16 was also hydrogen bonded to the phosphate (Table III; Figure III-7D).

In Arc repressor module M3, two of the three NHs, namely the NHs of Val33 and Asn34 on the N-terminus of the C-terminal α helix were free. Both NHs hydrogen bonded to the phosphodiester oxygen atoms of the same adenine. The NH of Ser35 was assumed to be free, yet it was hydrogen bonded to Oγ of Ser32. This side chain-main chain hydrogen bond is a typical Ncap (Richardson & Richardson, 1988) (Table III-3; Figure III-7E).

A common feature in all the five modules was the hydrogen bond between phosphodiester oxygen atoms of DNA and one of the NHs at the N-terminus of the C-terminal α helix. One more hydrogen bond donor was used for DNA binding out of the three NHs and/or side chains. In total, at least two

hydrogen bond donors were required. N-terminus of an α helix is positively charged, because of the dipole moment of the helix (Hol *et al.*, 1978; Warwiker & Watson, 1982). This dipole moment likely functions as an attractive force for DNA phosphates.

Three more common features amongst the pbHTH modules were found when examining the 3D structure of proteins without DNA and the amino acid sequences of pbHTH modules (Figure III-6B). First, the main chain NHs of the C-terminal α helix of pbHTH modules used for DNA phosphodiester-binding in complex with DNA formed no hydrogen bonds to other part of proteins, when the proteins did not bind DNA. Typically, there is an Ncap at the N-terminus of an α helix (Richardson & Richardson, 1988). The NH of the N-terminus of the C-terminal α helix in pbHTH modules was, however, not capped, and presumably formed hydrogen bonds to water molecules, as if waiting for DNA to present itself. Second, the size of a side chain of a residue located one residue before the C-terminal α helix is small. This is perhaps to avoid being a steric hindrance when the module binds to the DNA backbone. The side chain came close to DNA, since the main chain NH located next to the residue formed a hydrogen bond to a DNA backbone. Third, the pbHTH module has a number of arginines or lysines, residues with positively charged side chains within the module or juxtaposing ones. This seems to contribute to electrostatic interaction with DNA backbone (Figure III-7).

The repetitive appearance of the pbHTH module in DNA polymerase β, Oct-1 POU domain, 434 Cro and Arc repressor indicates that sequence non-specific binding of DNA is achieved by the protein module, and that the mode of sequence non-specific binding of DNA is not arbitrary; a common structure used for sequence non-specific binding was noted. The extent of these findings remains to be surveyed. Our study challenges the view regarding non-specific interaction of DNA-binding proteins in that there is no rule or pattern for DNA backbone contact in proteins (Pabo & Sauer, 1992).

Orientation of DNAs against pbHTH modules

Orientation of the DNA region against module M4 and that against module M7 in DNA polymerase β was similar. The orientation of the DNA region against Oct-1 POU domain module M2, 434 Cro protein module M2 and Arc repressor module M3 was also quite similar. However, the orientation of DNA to DNA polymerase β module M7 and that of Arc repressor M3 clearly differed (Figure III-8); the locations of NH contributing to hydrogen bonds in modules M4 and M7 of DNA polymerase β differ from those in module M2 of Oct-1 POU domain, in module M2 of 434 Cro and in module M3 of Arc repressor. In DNA polymerase β, the second and the third NHs of the N-terminus of the C-terminal α helix form hydrogen bonds with DNA, whereas in the Arc repressor, the first and the second NHs form hydrogen bonds with DNA. Two residues in an α helix are related by a rotation of about 100 degrees along the α helix axis. It is, therefore, geometrically calculated that difference in the orientation of the DNA axis in DNA polymerase β module M7 and Arc repressor module M3 is about 80 degrees (Figure III-9). The angle of DNA bound to DNA polymerase β and that bound to Arc repressor was actually about 80 degrees, when the modules were superimposed.

The difference in angle of bound DNA in DNA polymerase β modules M4 and M7, and Arc repressor module M3, 434 Cro module M2 or Oct-1 POU domain module M2 seems to relate to additional hydrogen bonds in pbHTH modules of DNA polymerase β. In modules M4 and M7, NHs of Gly64 and Gly105, respectively, form hydrogen bonds to DNA backbones. These hydrogen bonds did not exist in the Oct-1 POU domain module M2, 434 Cro module M2 nor Arc repressor module M3 (Figure III-7).

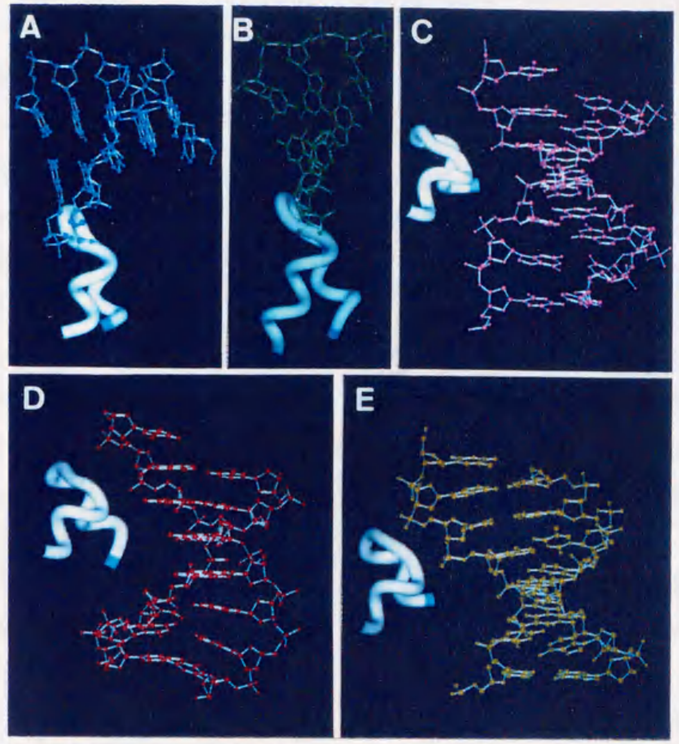

Fig III-8. DNA-binding mode against phosphate-binding HTH modules. (A) DNA polymerase β M4, (B) DNA polymerase β M7, (C) Oct-1 POU domain M2, (D) 434 Cro protein M2 and (E) Arc repressor M3. The pbHTH modules are viewed from the same side. DNAs for DNA polymerase β are located at the similar position to each other (A, B). DNAs for Oct-1 POU domain, 434 Cro and Arc repressor are also located in a similar position to one another (C, D, E), but different from those of DNA polymerase β.

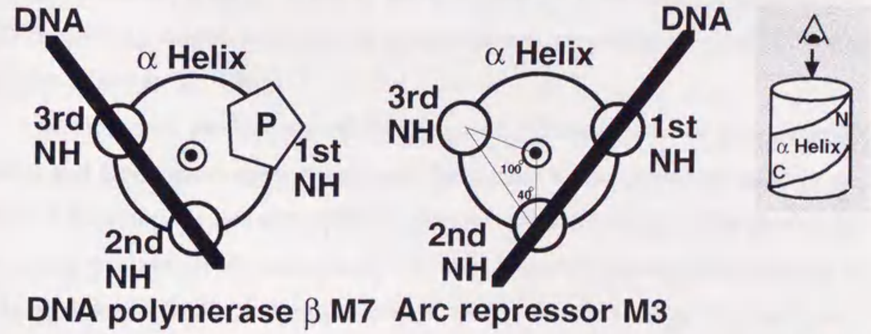

Fig III-9. N-terminal top view of α helix (see the inlet) from DNA polymerase β module M7 (left) and Arc repressor module M3 (right). A bound DNA is represented by a black bar. In DNA polymerase β module M7, the second and third protruding NHs form hydrogen bonds to phosphates of DNA, whereas in Arc repressor module M3, the first and second NHs form hydrogen bonds. The difference changes the orientation of DNA about 80 degrees.

III-5. Multiple functions of HTH modules

Types of HTH module: base-recognition, phosphate-binding and scaffold modules.

We reported that 434 Cro protein has three HTH modules, modules M2, M3 and M5 (Yura *et al.*, 1993). RMSD of these modules and against the ones found here are given in Table III-2. Module M2 of 434 Cro protein is a DNA phosphate-binding module shared by DNA polymerase β, Oct-1 POU domain and Arc repressor. Module M3 recognizes DNA bases. Module M5 seems to have no obvious functions (Yura *et al.*, 1993). The consequence is that there are at least three types of HTH module (Figure III-10).

The first HTH module is the newly identified pbHTH module. The backbone atoms on the N-terminus of the C-terminal α helix are used to form hydrogen bonds with the DNA backbone. The pbHTH module binds DNA, without sequence specificity. The second one is a base-recognition HTH (brHTH) module which is almost equivalent to the HTH motif. The C-terminal α helix is inserted into a major groove of DNA and side chains on the α helix hydrogen bond with DNA bases. The brHTH module binds DNA with sequence specificity. The third HTH module is a scaffold HTH (sfHTH) module with no obvious functions; it might possibly serve as a scaffold for other modules. The sfHTH module might function as a foundation to properly allocate other modules (Yura & Gō, 1995).

Motif-based predictions of DNA-bound 3D structures of transcription factors and DNA manipulation enzymes have been made (Ramakrishnan *et al.*, 1993). Characteristics of the pbHTH module are also likely to be useful for predicting protein-DNA interaction. Finding a module structurally similar to HTH modules in DNA-binding proteins of which the DNA-complex structure is unknown, and finding a module possessing features of the pbHTH module, but not of brHTH nor sfHTH modules, leads to the prediction that DNA may bind

to the module as the way it does in DNA polymerase β, Oct-1 POU domain, 434 Cro protein or Arc repressor.

Repetitive use of a structurally similar pbHTH module

A module similar in 3D structure and function to modules M4 and M7 of DNA polymerase β was found in Oct-1 POU domain, 434 Cro protein and Arc repressor. However, overall 3D structures of the four proteins differ (Figure III-5). This finding supports to the notion that the pbHTH module was shuffled into different proteins at the time of protein creation. One of the conceivable mechanisms here is exon shuffling (Gilbert, 1978; Long *et al.*, 1996). Of the four studied proteins, introns have been found in the genes of human DNA polymerase β (Chyan *et al.*, 1994) and its homolog, human terminal deoxynucleotidyl transferase (Riley *et al.*, 1988) and Oct-1. DNA polymerase β was divided into four subdomains (Pelletier *et al.*, 1994). Two out of twenty introns are located close to two of three subdomain boundaries (Figure III-3). The correspondence supports the widely accepted view that domains were indeed combined by exon shuffling (Doolittle, 1995). However, the other eighteen introns are also located in subdomains of DNA polymerase β (Figure III-3) and these introns tend to locate close to module boundaries, with some exceptions. Accumulation of protein 3D structures and their genomic structures resulted in a statistically significant correlation of module boundary and intron position (Gō & Noguti, 1995; de Souza *et al.*, 1996) A statistically significant correlation between module boundaries and intron positions in transcription factors was also noted (Yura *et al.*, in preparation). In DNA polymerase β, there was an intron on the N-terminus of pbHTH module M7, and in Oct-1, there were introns on the C-terminus of pbHTH module M2 in human, mouse and nematode genes (Figure III-3).

In 434 Cro protein, two phosphate-binding modules M2 and M4 sandwich the base-recognition module M3 (Figure III-5D). This sandwich

arrangement seems to be an effective mode to interact with DNA. Backbone of DNA seems to be held and pulled toward the protein by the two phosphate-binding modules. As a result, the base-recognition module located between the phosphate-binding modules seems to be inserted into a major groove of DNA and interacts with DNA bases. The division of labor in 434 Cro protein, namely base-recognition by module M3 and phosphate-binding by modules M2 and M4, could be an evolutionary remnant of the functional combination of modules. The combination of two different phosphate-binding modules and a base-recognition module as well as scaffold modules possibly created 434 Cro protein.

Module shuffling could shuffle a function to different proteins; in the present case, a function to bind phosphates. Isolated and incorporated modules are likely to have similar functions. Function of an isolated module was described by Yanagawa *et al* (1993). An isolated module of barnase, one of the RNases, cleaved RNA. Incorporation of a module into different combinations of modules may not affect the 3D structure of the module. Ikura *et al.* (1993) showed by NMR measurements that an isolated module had a tendency to take a 3D structure similar to its structure in intact protein. Conformation of a module is mostly determined by intra-module interaction (Noguti *et al.*, 1993). The pbHTH module may possibly have been incorporated into DNA polymerase β, POU-specific domain of Oct-1 POU domain, 434 Cro protein and Arc repressor.

Putative evolutionary relationship of pbHTH, brHTH and sfHTH modules

The pbHTH module exists in DNA polymerase β and Oct-1 POU domain of eukaryotes, and 434 Cro protein and Arc repressor of prokaryotes. The brHTH module as an HTH motif exists in eukaryotes and prokaryotes (Luisi, 1995). Therefore, pbHTH and brHTH modules probably existed before the divergence of eukaryotes and prokaryotes.

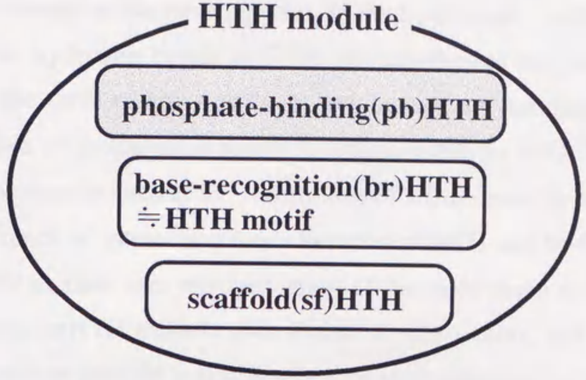

Fig III-10. Classification of HTH modules based on their functions. There are at least three functions in HTH modules. HTH modules are functionally grouped into pbHTH, brHTH, and sfHTH modules.

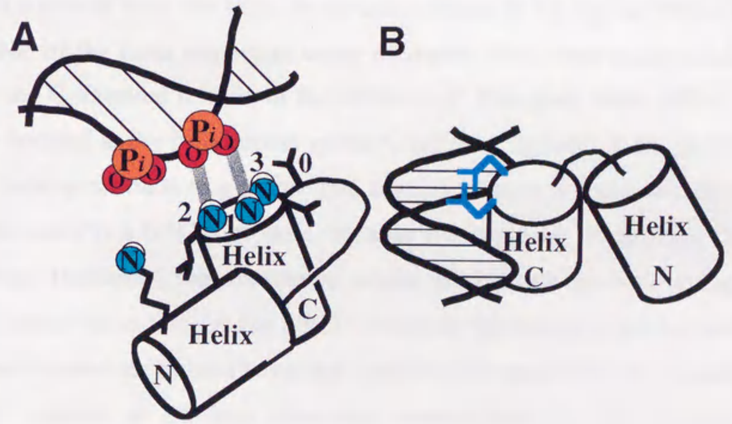

Fig III-11. Difference in binding mode of (A) pbHTH and (B) brHTH of prokaryote to DNA. Side chains of residues that form hydrogen bonds or seemingly strong electrostatic interaction are drawn. C-terminal α helix of pbHTH module is almost perpendicular to the DNA backbone, whereas that of brHTH is inserted into a major groove.

The brHTH and pbHTH modules could be evolutionary related, for they use residues located at the same position on the C-terminal α helix. The pbHTH module forms hydrogen bonds to DNA phosphodiester oxygens by using at least one of the three residues at the N-terminus of a C-terminal α helix. The brHTH module of prokaryotes forms hydrogen bonds to DNA bases by using one or two residues located at the N-terminus of a C-terminal α helix (Table III-4). The difference of interaction mode between pbHTH and brHTH modules is that 1) pbHTH module uses nitrogen atoms of the main chain to form hydrogen bonds, whereas brHTH module uses atoms of side chains, and that 2) the C-terminal α helix of pbHTH module is located perpendicular to the backbone of DNA, whereas the C-terminal α helix of brHTH module is inserted into a major groove of DNA (Figure III-11).

The HTH module of the *trp* repressor has characteristics of both brHTH and pbHTH modules. DNA complex of *trp* repressor revealed that water-mediated contacts were the most important contacts to recognize DNA (Sigler, 1992). One of the most important water mediated DNA-base recognitions was found at the C-terminal α helix of the HTH motif. The main chain NH of Ala80 which is located at the N-terminus of the C-terminal α helix in the HTH motif formed hydrogen bonds to guanine and adenine bases via the water molecule. This HTH motif is a brHTH module, because it recognizes DNA bases (Yura *et al.*, 1993). However, the manner in which it realized the base-recognition function is similar to that for the pbHTH module, because it used the backbone NH of the N-terminus of the C-terminal α helix of the module. We consider that the HTH module in the *trp* repressor corresponds to the evolutionary intermediate between pbHTH and brHTH modules (Figure III-12).

Which type of HTH module appeared first? Phosphate is a fundamental component of DNA/RNA, but not of proteins. To bind specifically to DNA/RNA, but not to proteins, finding phosphorus atoms is considered to be the most effective. Binding to phosphates could therefore be a fundamental to

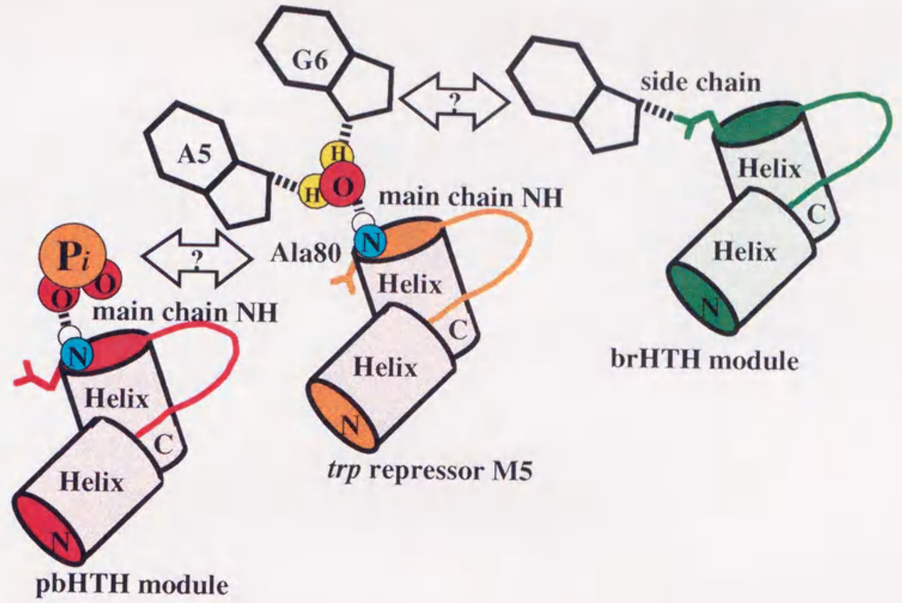

Fig III-12. Schematic drawing to emphasize characteristics of the brHTH module in *trp* repressor. The brHTH module in *trp* repressor forms hydrogen bonds to DNA bases, using protein backbone NH via the water molecule. The manner in which a protein backbone contributes to hydrogen bonds between protein and DNA is similar to the case of the pbHTH module.

interact with DNA/RNA. During the course of pre-biological evolution, the pbHTH modules might appear first, since the pbHTH module binds to phosphates of DNA. Some pbHTH modules might acquire the potential to bind to DNA bases, using a water molecule as in case of the *trp* repressor. Those pbHTH modules could become brHTH modules by using their side chains instead of a water molecule to interact with the nucleotide. That would be the emergence of the base specifically interacting HTH module, brHTH module. The possibility that the brHTH module appeared first is not ruled out. The sfHTH modules might emerge as a result of recruitment from pbHTH or brHTH module. The sfHTH modules could at first interact with DNA/RNA, but possibly lost the ability and specialized as a scaffold. The sfHTH module was probably needed to stabilize a globular domain (Figure III-13).

DNA is apparently a late comer in evolution, compared to RNA (Joyce & Orgel, 1993). The first pbHTH and brHTH modules could interact with RNA. There is no direct evidence of HTH module-RNA interaction, but several lines of evidence, such as homeodomain-RNA interaction (Dubnau & Struhl, 1996), HTH motif-like structures in L11 ribosomal protein (Hinck *et al.*, 1997) and glutamyl-tRNA synthetase (Nureki *et al.*, 1995) suggest the possibility of RNA-HTH module interaction.

Table III-4. Hydrogen bonds between brHTH modules and DNA-bases

Protein Name	1st res.	2nd res.
catabolite gene activator	-	E181(Oε2) - C_5(N4)
hin recombinase	-	S174(Oγ) - A_{10}(N7)
λ repressor	Q44(Oε1) - A_4(N6)	S45(Oγ) - G_{16}(N7)
purine repressor	-	T16(Oγ1) - A_6(N6)
434 Cro	Q28(Oε1) - A_5(N6)	Q29(Nε2) - G_{16}(O6)
lac repressor	Y17(Oη) - T_4(O4)	Q18(Oε1) - C_5(N4)

"1st res." and "2nd res." indicate the first and the second residues of the C-terminal α helix of brHTH module, respectively. In each column, the left side is a protein residue(atom) and the right side is a DNA nucleotide(atom). Atoms in parenthesis are a hydrogen donor or an acceptor. Protein-DNA hydrogen bonds were calculated based on the PDB entry of 1CGP (Schultz *et al.*, 1991) for catabolite gene activator, 1HCR (Feng *et al.*, 1994) for hin recombinase, 1LRD (Jordan & Pabo, 1988) for λ repressor, 1PNR (Schumacher *et al.*, 1994) for purine repressor, 3CRO (Mondragon & Harrison, 1991) for 434 Cro protein, and 1LCC (Chuprina *et al.* 1993) for lac repressor. Hydrogen bonds were calculated by the method of Baker and Hubbard (1984).

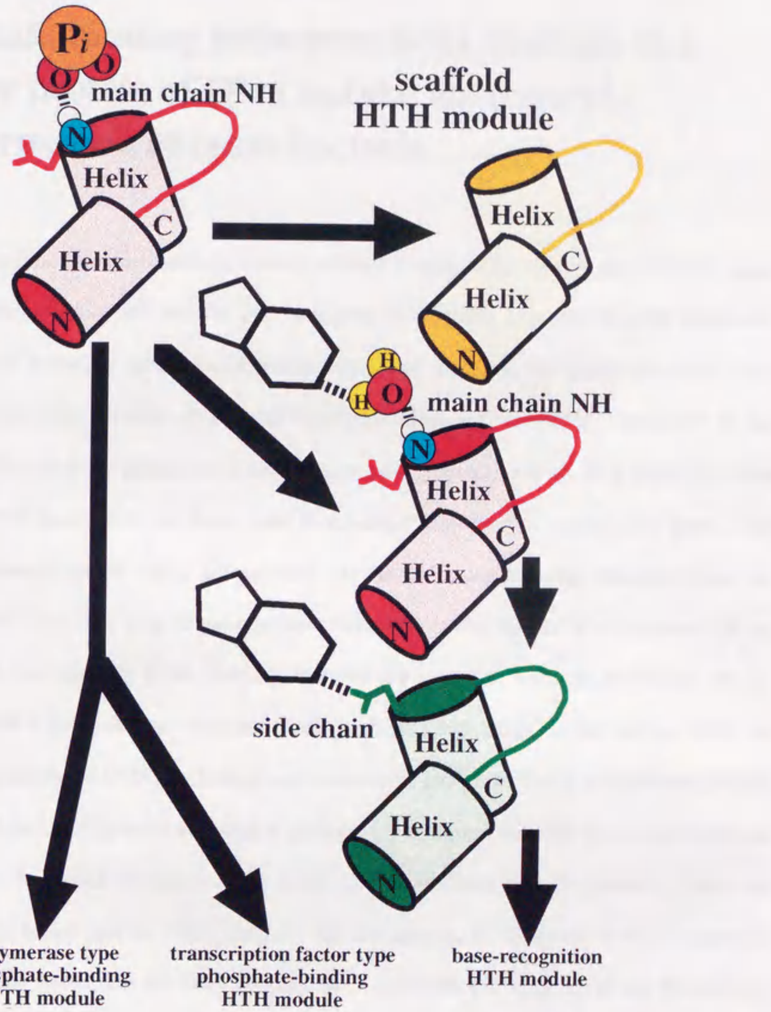

Fig III-13. Putative functional divergence of HTH modules. A red module is the pbHTH module, a green module is the brHTH module, and a yellow one is the sfHTH module. Initial pbHTH module hydrogen bonds with phosphodiester oxygen by a main chain NH. A brHTH module might be modified out of the pbHTH module by changing the hydrogen acceptor from a phosphate to a base with help from a water molecule. A side chain probably took on the roles of NH and water molecule, at the last stage. An sfHTH module emerged as a result of recruitment of the pbHTH module or brHTH module for a foundation for other functional modules to make fused modules, as a stable globular domain.

Chapter IV.
Phosphate-binding helix-turn-helix modules in a putative protein of DNA uptake for natural transformation of cyanobacteria

Abstract: Genetic transformation is widely utilized in molecular biology as a tool for gene cloning in *Escherichia coli* and for gene mapping in *Bacillus subtilis*. Several strains of eubacteria can naturally take up exogenous DNA and integrate the DNA into their own genomes. Molecular details of natural transformation are, however, remained to be elucidated. The complete genome of a cyanobacterium, *Synechocystis* sp. PCC6803, has been sequenced. This bacterium has been used to examine functions of a particular gene. The genome is considered to carry information on natural transformable characteristics of *Synechocystis*. The first step in genetic transformation is the uptake of exogenous DNA. Proteins with non-specific DNA binding features are required, because specificity in the exogenous DNA has not been evident. Such proteins have modules interacting with the phosphate backbone of DNA, including helix-turn-helix modules. Using a consensus pattern of the phosphate-binding helix-turn-helix module, we searched through the genome data of *Synechocystis* for genes or open reading frame (ORF) products with the pattern in primary structures. We found that an ORF, slr0197, has the pattern, in duplicate at the C-terminal region. We also found that the ORF product has a hydrophobic segment at the N-terminal region, which is followed by two-fold internal repeat of endonuclease domain. Based on these observations, we propose a model for the initial stage of genetic transformation. This is apparently the first report on molecular mechanisms of natural transformation.

IV-1. *Synechocystis* and natural transformation

Since the discovery of DNA as genetic material by transformation of *Pneumococcus* (Griffith, 1928), studies on genetic transformation have continued. Genetic transformation is a process that bacterial cell uptakes exogenous DNA, incorporates it and expresses the gained trait (Dreisenkelmann, 1994; Solomon & Grossman, 1996). The transformation process is divided into separate steps as, binding to DNA, processing and uptaking of DNA and integration of DNA into the chromosome (Dreisenkelmann, 1994). Some cyanobacteria are able to take up DNA under normal physiological conditions (Porter, 1986). We consider that the complete genome (Kaneko *et al.*, 1996) of a cyanobacterium, *Synechocystis* sp. PCC6803 carries information on the natural transformable characteristics of *Synechocystis*. Conventional annotation methods specified functions for about 50% of the genes and ORF products of the total genome in *Synechocystis* sp. strain PCC6803 and functions for the remaining are unknown (Kaneko & Tabata, 1997). Such being the case, we used a new approach to search for genes or ORF products dealing with natural transformation of *Synechocystis*.

The initial step of DNA uptake for transformation involves the capture of exogenous DNA. *Synechocystis* sp. strain PCC6803 can virtually take up any sequence of DNA (Grigorieva & Shestakov, 1982). Non-specific interactions between DNA and a protein on a transcription factor-DNA complex were found to be dominated by interactions between DNA phosphate backbones and proteins (Albright *et al.*, 1998). Studies of three-dimensional (3D) structures of transcription factors and DNA manipulation enzymes revealed similar patterns in sequence/structure widely used for non-specific interactions between protein and DNA (Doherty *et al.*, 1996). The DNA phosphate-binding helix-turn-helix

(pbHTH) module was found in three transcription factors and a DNA repair enzyme (Yura *et al.*, 1999). A module, defined as a compact structure within a globular domain (Gō, 1981), on average consists of about 15 contiguous amino acid residues. The correlation of the module organization of a protein with exon/intron structure of the gene (Gō, 1983; Gō, 1985; Gō & Noguti, 1995; de Souza *et al.*, 1996) and the distribution of similar modules in different proteins (Yura *et al.*, 1993) indicate that a module may be a unit of protein function. Locating pbHTH module in the whole genome of *Synechocystis* will pinpoint proteins that bind DNA non-specifically, including a candidate protein for DNA uptake.

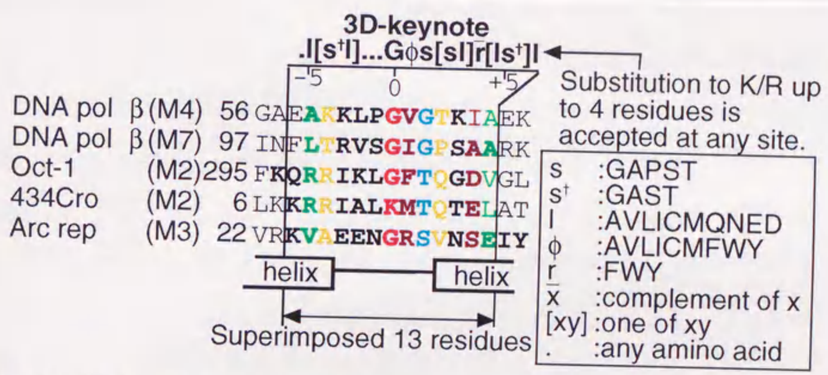

Fig IV-1. 3D-keynote of the pbHTH module. Residues in bold font make up the module. Colored residues form the core of the module, except for the red and blue residues; red residues are for turn structures and blue residues locate close to DNA backbones. In 3D-keynote, s is short for small side chains, s† for small side chain without proline, l for aliphatic, ϕ for hydrophobic, r for aromatic, upper bar for complement, square bracket for choice and dot for any residue. Protein Data Bank code for each protein is 1bpy for DNA polymerase β, 1oct for Oct-1, 3cro for 434 Cro protein and 1par for Arc repressor.

IV-2. 3D-keynote

3D-keynote of pbHTH

Considering structural requirements for interactions with the phosphate backbone of DNA, together with the residue feature at each site obtained from the alignment of pbHTH module, we constructed a distinct sequence pattern for the module, which we named 3D-keynote in order to distinguish it from sequence motif simply obtained from residue conservation. Actually, the residue level conservation at each site of pbHTH module was too weak to construct a sequence motif (Figure IV-1). Interactions with the phosphate backbone of DNA require rules on amino acid residues at positions +2 and +4. A phosphate interaction site of pbHTH modules is a backbone amino hydrogen on the N-terminus of the C-terminal α helix. A phosphorous oxygen of DNA and the hydrogen form a hydrogen bond (Yura *et al.*, 1999). This interaction limits the size of side chain at position +2 as small, because a DNA backbone and the residue locate close by. The interaction also limits the size of side chain at position +4 as non-bulky. A structural feature of pbHTH modules is represented by side chain contacts among residues at positions -5, -4, +1, +3, +5 and +6. They form a core structure of the module. Interactions of those side chains are expressed by pairwise side chain interactions by finding a pair of the residues in extensive contact. In Figure IV-1, the residues in contact are shown in the same color. The pairs of residue were conserved to maintain the volume of the core. The residue at position 0 (red) is required for turn structure formation. Proline cannot reside in an α helix except for the N-terminus, since proline disrupts a main chain hydrogen bond. To interact with DNA phosphate electrostatically, one to four acidic residues are required at any site within the module (Yura *et*

al., 1999) (Figure IV-1). The rules at each amino acid position of pbHTH module deduced from the functional and structural requirements were expressed as 3D-keynote by a code representing one of the groups of amino acid residues. In Figure IV-1, the groups of amino acid residues were represented by a single letter with superscript when necessary. The code s includes weak polar residues (Gō & Miyazawa, 1980). The code $s^{\dagger}$ includes weak polar residues except proline; proline appears at limited sites in an α helix as mentioned above. The code l includes residues with an aliphatic side chain; a polar residue in this group uses aliphatic part of the side chain to form a core structure of the module. The code ϕ includes hydrophobic residues and r includes aromatic residues (Dayhoff *et al.*, 1978).

Finding sequences that match the 3D-keynote

In the whole genome of *Synechocystis* sp. PCC6803, candidate sequences were initially selected with a similarity search carried out by BLAST (Altschul *et al.*, 1990) with default value implemented in Cyanobase (Kaneko & Tabata, 1997). Five sequences in Figure IV-1 were used as a query sequence. Candidate sequences were filtered by 3D-keynote.

ORF slr0197, ComEA

In total genome of *Synechocystis*, the 3D-keynote of pbHTH module was found in ORF slr0197, DNA helicase II and excinucleases ABC subunit C. A product of ORF slr0197 was found to have two pbHTH modules (Figure IV-2A) including the helix-hairpin-helix (HhH) motif reported by Doherty *et al* (1996). ORF slr0197 was annotated as ComEA, based on partial similarity to *Bacillus subtilis* ComEA (Kaneko & Tabata, 1997). The product of slr0197 is 553 amino

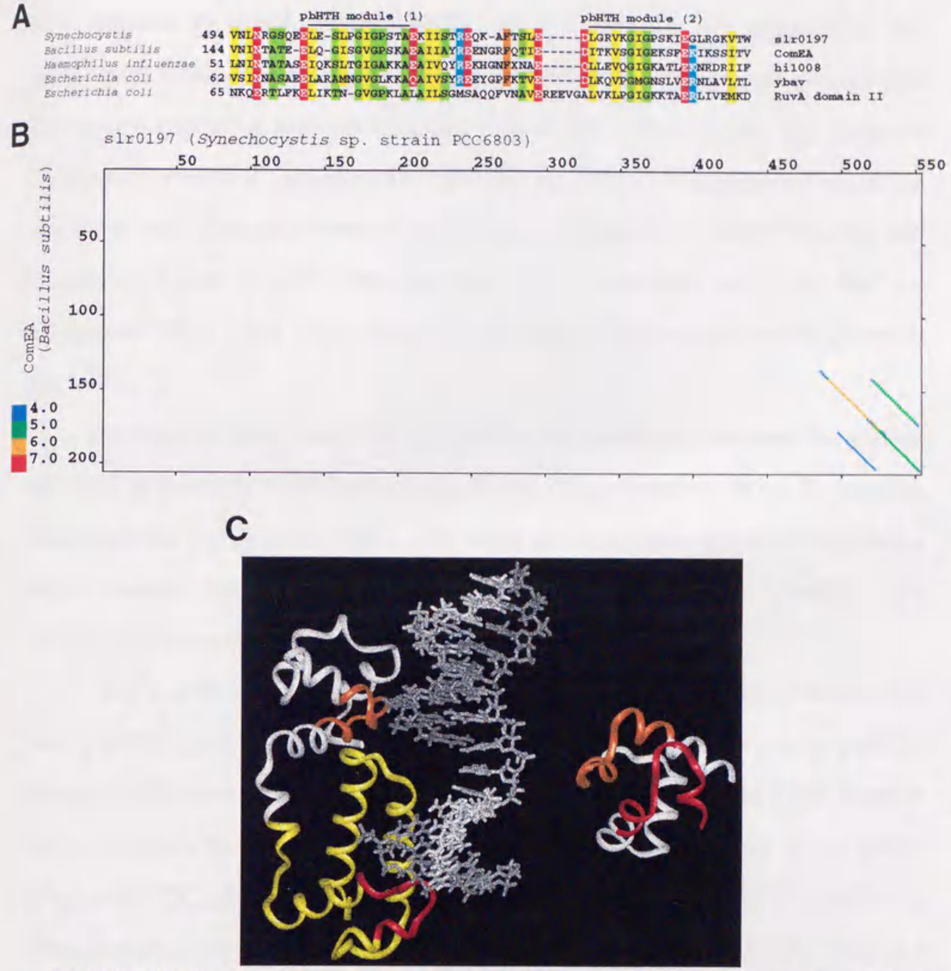

Fig IV-2. (A) Amino acid sequences that match 3D-keynote in Fig IV-1. in duplicate. Similar amno acid residues are coloured. Accession numbers of the sequences; DNA polymerase β is 1bpy (PDB), slr0197 is SLR0197 (Cyanobase), ComEA is L15202 (SwissProt), hi1008 is HI1008 (Tiger), Ybav is P77415 (SwissProt) and RuvA is 1hjp (PDB). (B) Harr plot of Synechocystis slr0197 and B. subtilis ComEA. Only C-terminal region is similar. A colour of dot corresponds to significance of similarity expressed by Z score (McLachlan, 1971). (C) The 3D structure of DNA polymerase β N-terminal and Fingers domains (left:1bpy) (Sawaya et al., 1997) and RuvA domain II (right:1hjp) (Nishino et al., 1998) that have two pbHTH modules. Red and orange modules are pbHTH modules. DNA locates between two pbHTH modules of RuvA (Hargreaves et al., 1998). In Synechocystis slr0197, DNA is likely to bind as in the case of DNA polymerase β.

acid residues in length, while ComEA from *B. subtilis* is composed of 205 amino acid residues. The sequence similarity between them was restricted to the C-terminal region of each protein, and size of the similar region was about 60 amino acid residues, including two pbHTH modules. No significant sequence similarity was observed between the N-terminal regions of slr0197 product and ComEA (Figure IV-2B). The protein from *B. subtilis* serves to bind an exogenous DNA when competence for genomic transformation arises (Hahn *et al.*, 1993).

In *Synechocystis*, only slr0197 has the sequence that matches the pattern of pbHTH module 3D-keynote in duplicate. The sequences from *B. subtilis*, *Haemophilus influenzae*, and *E. coli* were obtained by a sequence similarity search against the sequence of approximately 60 residues in *Synechocystis* slr0197 that contains two pbHTH modules (Figure IV-2A).

The C-terminal domain of slr0197 would bind DNA non-specifically with two pbHTH modules. DNA polymerase β and RuvA have two pbHTH modules/HhH motifs and interact with the phosphate backbone of DNA through these modules(Doherty *et al.*, 1996; Nishino *et al.*, 1998; Yura *et al.*, 1999) (Figure IV-2C). The 3D structures support the notion that pbHTH modules in *Synechocystis* and other transformable eubacteria are bound to DNA. Thus, we termed the C-terminal domain DNA-binding domain.

IV-3. Endonuclease domain with HKD motif

In the N-terminal 500 residues of slr0197, a two-fold internal duplication was identified; these duplicated domains are similar in amino acid sequence not only to each other, but also to the endonucleases characterized by the

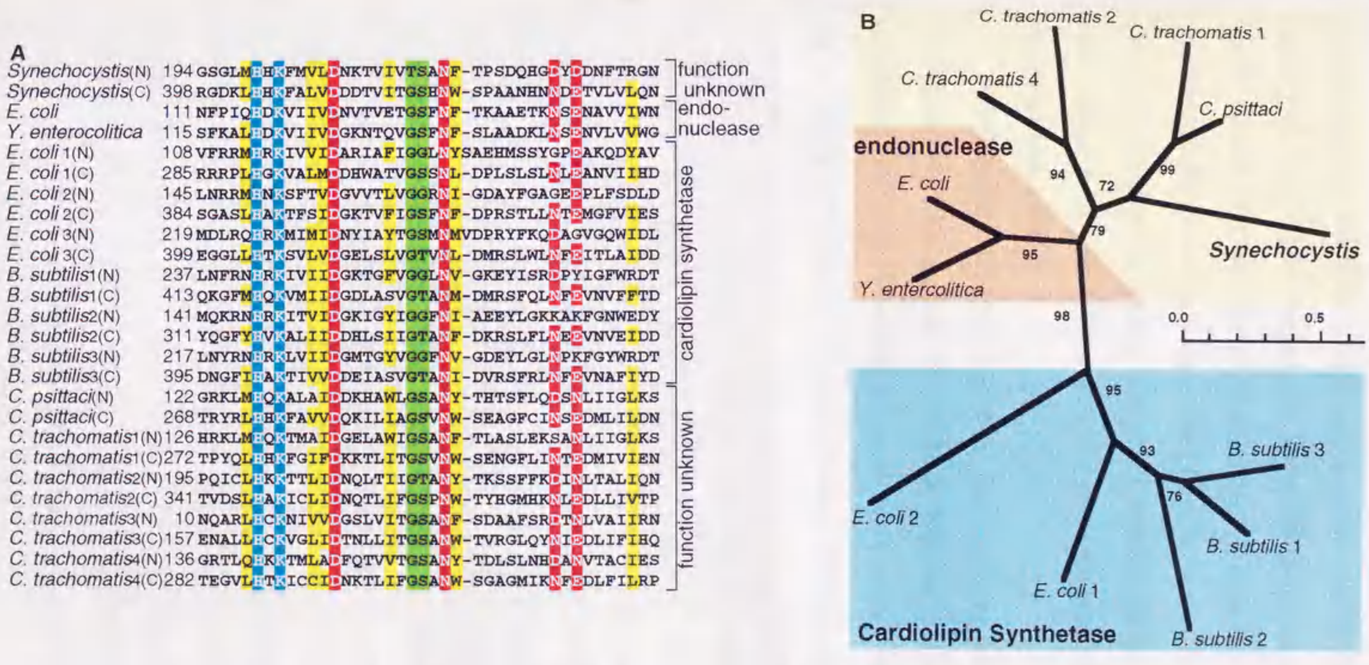

Fig IV-3. Alignment and phylogenetic tree of domains with HKD motif. (A) Sequence alignment of central regions of *Synechocystis* slr0197 against HKD motifs in *E. coli*, *Yersinia enterocolitica*, *B. subtilis*, *Chlamydia psittaci* and *C. trachomatis*. Histidine, lysine, aspartic acid and aspargine are identically conserved and considered to form active sites. (N) and (C) after the name of species indicate N-terminal and C-terminal motif, respectively. Accession numbers are, from the top of the alignment, SLR0197(Cyanobase), SLR0197(Cyanobase), Q46707 (SwissProt), O07482 (SwissProt), AE000181 (GenBank), AE000181 (GenBank), AE000206 (GenBank), AE000206 (GenBank), AE000223 (GenBank), AE000223 (GenBank), P45860 (SwissProt), P45860 (SwissProt), P45865 (SwissProt), P45865 (SwissProt), BG12483 (SwissProt), BG12483 (SwissProt), O34024 (SwissProt), O34024 (SwissProt), G3328479 (SwissProt), G3328479 (SwissProt), G3328559 (SwissProt), G3328559 (SwissProt), G3328561 (SwissProt), G3328561 (SwissProt), G3328556 (SwissProt) and G3328556 (SwissProt). (B) An unrooted tree of conserved regions of the domains with HKD motifs (85 residues). The tree was drawn by the method of neighbor joining (Saitou & Nei, 1987). Names of species and number correspond to those in (A). Only sequences around C-terminal motif were used. Number at each branch is a bootstrap probability calculated on 1,000 re-sampling. The obtained tree topologies of neighbor joining and maximum likelihood methods were virtually the same (data not shown).

conservation of histidine, lysine and aspartic acid (Figure IV-3A). A similar result was obtained by Ponting and Kerr (1996) and Koonin (1996). The HKD motif, named after the conserved residues, has been found in *Escherichia coli nuc* endonuclease and cardiolipin synthetase from several eubacteria. In Figure IV-3A, two sequences from *E. coli* and *Y. enterocolitica* are endonucleases and others from *E. coli* and *B. subtilis* are cardiolipin synthetases. Functions of others are unknown.

Phylogenetic analysis was done on sequences with HKD motif. A tree was drawn by the method of neighbor joining (Saitou & Nei, 1987) and maximum likelihood (Strimmer & von Haeseler, 1996). Distance was calculated by a similarity-distance method (Fukami-Kobayashi, 1994). Only well conserved parts in the amino acid sequences were used for tree making. A tree topology was evaluated by a bootstrap probability calculated on 1,000 re-sampling. The analysis indicated that the region in *Synechocystis* slr0197 is closer to ones from endonucleases than ones from cardiolipin synthetases (Figure IV-3B). Branching of the tree corresponds to functional group of the proteins. HKD motif in *Synechocystis* and the motif in *Chlamydia* are grouped with endonucleases. HKD motifs in cardiolipin synthetases from different organisms form a distinct group from that of endonucleases.

E. coli nuc endonuclease, encoded on a eubacterial drug resistance IncN plasmid (Pohlman *et al.*, 1993), is a periplasmically localized, EDTA resistant endonuclease that degrades single- and double-stranded DNA with virtually no specificity. The enzyme was considered to play some role in conjugation, because it appeared to be co-translated with a cluster of genes for conjugal transfer (Pohlman *et al.*, 1993). An active site was probably formed by conserved residues in the motif, namely a histidine, a lysine, an aspartic acid

and an aspargine (Pohlman *et al.*, 1993; Ponting & Kerr, 1996; Koonin, 1996).

IV-4. Domain organization and function of ORF slr0197

Hydropathy analysis was done on slr0197 to predict membrane spanning regions. The method of Kyte and Doolittle (1982) was used. A window length for an index of each amino acid residues was set to seven centered on the residue. The sequence of N-terminal 30 residues of *Synechocystis* slr0197 turned out to be rich in hydrophobic residues (Figure IV-4A). The hydrophobicity indicates that *Synechocystis* slr0197 is a transmembrane protein with one membrane-spanning region at its N-terminus. We interpret the product of ORF slr0197 to be a periplasmic membrane protein, because 1) N-terminal 30 amino acid sequence was predicted to be a membrane spanning region (Figure IV-4A), 2) the two central domains are similar to *E. coli nuc* endonuclease which is located in the periplasm (Figure IV-3B) and 3) the C-terminal domain is similar to part of *B. subtilis* ComEA which is located outside the membrane (Inamine & Dubnau, 1995).

The domain organization indicates that slr0197 is a protein for DNA uptake at the process of gene transformation for cyanobacteria, because 1) two pbHTH modules exist in the C-terminal domain that binds DNA non-specifically as in DNA polymerase β or RuvA (Figure IV-2C), 2) the C-terminal domain of slr0197 is similar to the C-terminal domain of *B. subtilis* ComEA which functions as a DNA-binding domain for DNA uptake, 3) two central domains, each with the HKD motif, in slr0197 are similar to *E. coli nuc* endonuclease which is likely to function for DNA uptake and/or transport (Pohlman *et al.*, 1993), 4) in the total genome of *Synechocystis* sp. PCC6803,

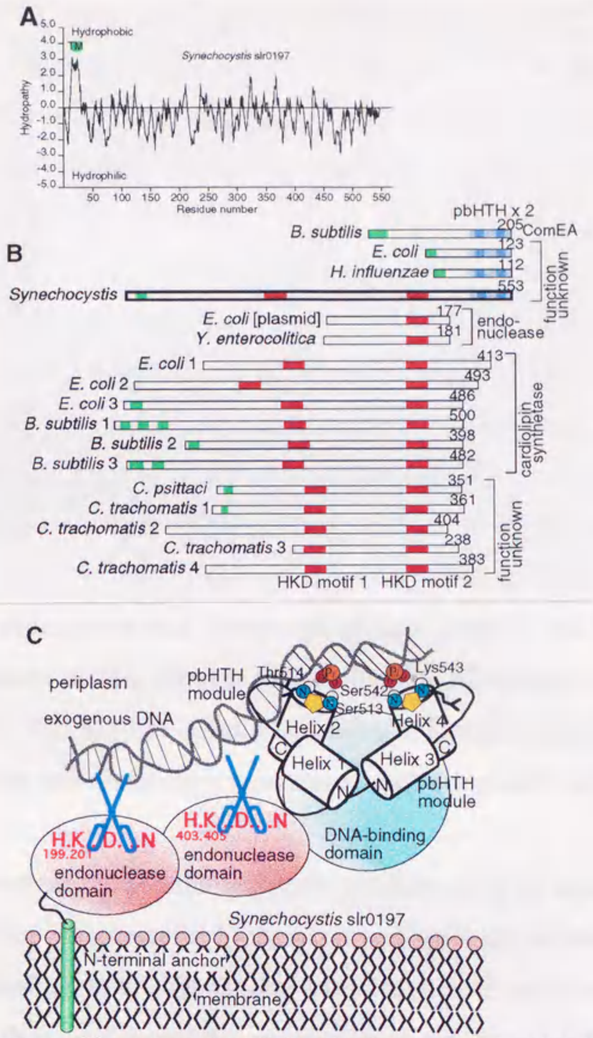

Fig IV-4. Overall structure of *Synechocystis* slr0197. (A) Hydropathy plot of slr0197. A highly hydrophobic region is found at the N-terminus. (B) Comparison of structures among *Synechocystis* slr0197 and proteins from other organisms with two pbHTH modules and/or HKD motifs. A green box is a putative transmembrane region, a red box is HKD motif shown in Fig IV-3A. and a blue box has two pbHTH modules (deep blue). Note that only *Synechocystis* slr0197 has the three colored boxes. Accession numbers are, from the top, L15202 (SwissProt), P77415 (SwissProt), HI1008 (Tiger), SLR0197 (Cyanobase), Q46707 (SwissProt), O07482 (SwissProt), AE000181 (GenBank), AE000206 (GenBank), AE000223 (GenBank), P45860 (SwissProt), P45865 (SwissProt), BG12483 (SwissProt), O34024 (SwissProt), G3328479 (SwissProt), G3328559 (SwissProt), G3328561 (SwissProt) and G3328556 (SwissProt). (C) Putative structure and functional sites of slr0197 of *Synechocystis* sp. strain PCC6803. N-terminal residues are assumed to be buried in an inner membrane. Putative active sites of two endonucleases with HKD motif are shown. In the C-terminal domain, two pbHTH modules would interact with DNA.

ORF slr0197 is the only amino acid sequence that matches the 3D-keynote and that has the HKD motif and 5) no other proteins encoded in the total genome of *Synechocystis* sp. PCC6803 have domains known to be involved in the DNA uptake in other organisms (Figure IV-4B).

The relationship between the function and the putative domain structure of ORF slr0197 product that we propose is shown schematically (Figure IV-4C); the first step of transformation, non-specific binding to DNA, is performed by the C-terminal domain and the second step of transformation, DNA processing, is performed by central endonuclease domains. In the C-terminal DNA-binding domain, the amino hydrogens of Ser513 and Thr514 in the first pbHTH module and those of Ser542 and Lys543 in the second pbHTH module are candidates for hydrogen donors used in phosphates binding. The central non-specific DNA endonuclease domains have active sites on His199, Lys201, Asp206 and Asn216 for the first domain and His403, Lys405, Asp410 and Asn420 for the second domain, because these residues are completely conserved (Figure IV-3A).

One cannot exclude the possibility that the proteins for DNA uptake are encoded in plasmids of cyanobacteria. In the sequenced plasmids, however, we found no proteins dealing with DNA uptake. In addition, both sequenced and unsequenced plasmids from different cyanobacteria with natural transformation characteristics are highly diverged in size, which means that common proteins in different cyanobacteria for DNA uptake are probably not encoded in such a plasmid.

IV-5. Molecular mechanisms of transformation in eubacteria

Synechocystis, *B. subtilis* and *E. coli* are transformed by taking up exogenous DNA (Porter, 1986). The total genome of those organisms has been sequenced and proteins with similar domain organization for DNA uptake were found (Kaneko *et al.*, 1996; Kunst *et al.* 1997; Blattner *et al.*, 1997). Each organism uses a different combination of proteins for related mechanisms (Figure IV-5). In the total genome of *B. subtilis, E. coli* and *Chlamydia trachomatis*, there is no single protein that has the DNA-binding domain similar to the one in slr0197 and HKD motif, at the same time. *Synechocystis* slr0197 is a unique protein with two endonuclease domains and a DNA-binding domain among eubacteria with known total genomes (Figure IV-4B). The fusion of the three domains would ensure efficient natural transformation. Transport of DNA into the cytoplasm and recombination no doubt involve other unidentified proteins. *B. subtilis* has ComEA, a DNA-binding protein for transformation (Inamine & Dubnau, 1995) and *E.coli* has a protein, both of which have a domain similar to the DNA-binding domain of *Synechocystis* slr0197 (Figure IV-2B), but these proteins do not have HKD motifs; *B.subtilis*, *E.coli* and *Synechocystis* have different proteins for processing of exogenous DNA. *E. coli* has an endonuclease with the HKD motif on a plasmid, but in the genome of *B. subtilis*, all of the HKD motif containing proteins are apparently used for cardiolipin synthetases. Therefore, an unknown endonuclease is likely to be used for exogenous DNA cleavage in *B. subtilis*. *C. trachomatis* is not known to have the potential to take up exogenous DNA nor the potential for natural transformation. However, total genome sequencing of the organism showed that 35 proteins more closely resemble eukaryotic homologs than do the eubacterial counterparts (Stephens *et al.*, 1998), which means that *C. trachomatis* has a

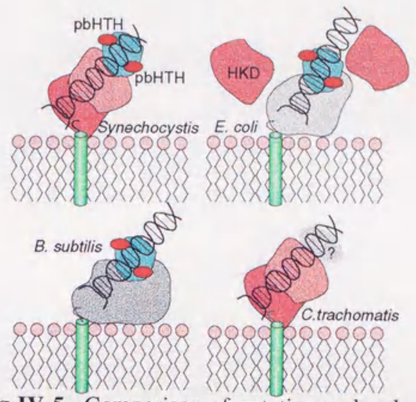

Fig IV-5. Comparison of putative molecules for DNA uptake in four organisms. Domain organizations of the molecules in different organisms are not the same. Green is a membrane spanning region, red endonucleases with HKD motif and blue DNA-binding domain with two pbHTH modules in orange.

system to take up exogenous DNA. *C. trachomatis* has four proteins with the HKD motif; the amino acid sequence of one of the proteins is clustered with the *nuc* endonuclease, but not with cardiolipin synthetase in a phylogenetic tree (Figure IV-3B) and has a putative transmembrane region (Figure IV-4B). This finding indicates that the protein is part of the exogenous DNA uptake machinery. The protein does not have a DNA-binding domain, so that a separate DNA-binding protein probably assists in exogenous DNA capture. Our observations suggest that several eubacteria have independently acquired genetic transformation machinery and that parts of the machinery are similar. The existence of similar domains for DNA uptake is limited to eubacteria as no similar sequences to ComEA nor to the HKD motif were detected in total genomes of several archaea and *Saccharomyces cerevisiae*.

We have proposed herein a candidate protein involved in natural

transformation of cyanobacteria. This protein can serve as one of the targets for mutations and a knockout to disclose mechanism and evolution of the exogenous DNA uptake of bacteria under normal physiological conditions.

Chapter V.
Conclusion and future direction

Transcription factors and DNA manipulation enzymes are decomposed into several domains. It has been conceived that DNA-binding domains take the role of interactions with DNA as a whole. The domains have been named after their sequence/structure motifs. Advance in structure determination methods unveiled atomic details of DNA-binding domains and showed that 1) not a whole domain, but a small part of the domain is actually in contact with DNA, and 2) residues in contact with DNA are not limited to residues in sequence/structure motifs. The domains need to be decomposed into certain smaller segments.

Decomposition of DNA-binding domains into modules, a compact unit of a contiguous segment, displayed that helix-turn-helix motif, one of the DNA-binding motifs, corresponded to a single module (Chapter II). The decomposition further showed that DNA-binding domains are built up by three types of modules, namely base-recognition, phosphate-binding and scaffold modules. Contact between DNA molecule and proteins is found to be subdivided by the modules (Chapter III). Division of functions in the domains by modules supports a model that the domain was evolved by module fusion. A different combination of modules could easily build domains that have different functions.

A base-recognition helix-turn-helix module was found in transcription factors of eubacteria (Chapter II). Three-dimensional structures of those domains are different from one another as a whole, but the module has a similar structure/sequence. A phosphate-binding helix-turn-helix module was found in transcription factors of eubacteria and eukaryotes and DNA repair enzymes of

eukaryotes. Numerous proteins, including proteins that seemingly do not interact with DNA/RNA have a scaffold helix-turn-helix module that works as a foundation for other functional modules (Yura & Go, 1995). These facts are strong pieces of evidence to indicate that a domain was evolved by combination of modules. The finding that the phosphate-binding helix-turn-helix module was found in the extracellular domain of ComEA was striking in a sense that the module was likely to be used widely in a situation wherever a protein binds any DNA (Chapter IV).

So far, evidence for module shuffling is found in modules with the helix-turn-helix structure. An urgent question to be answered is other cases of module shufflings. RNA-protein interactions are the best targets for addressing the question. RNA-binding proteins are supposed to be vestiges of RNA world and the interactions established at that stage remain in them. A present protein 3D structure database contains plenty of RNA-protein complex structures (Mattaj & Nagai, 1995). Glutamyl-tRNA synthetase, one of the RNA-binding proteins, was shown to have modules with helix-turn-helix structure. The module does interact with anticodon loop of glutamyl-tRNA (Tateno *et al.*, 1995).

The whole works here uncovered the modularity in structure/function of transcription factors and DNA manipulation enzymes. Functionally/ structurally similar modules were found in different DNA-binding proteins. These findings suggest a new method to predict a DNA-complex structure of transcription factors and DNA manipulation enzymes of which three-dimensional structure is known, but of which DNA-complex structure is unknown. Finding a module known to interact with DNA bases or phosphates in the target protein will delimit an orientation of the protein to a DNA molecule. There will be a method to predict a function of a module from its three-dimensional structure.

A finding that module is a functional unit can and will be applied to

genome sequence analyses. Emergence of genome sequences from several organisms unveiled that, at best, 50% of ORFs have reasonable sequence similarity to known proteins. The remaining 50% of ORFs are assigned as function unknown (Koonin & Galperin, 1997). The present method of function identification is based on the notion that proteins evolved by domain fusion. Therefore, similarity searches are limited to the size of domain. Search of a short motif such as P-loop (Walker *et al.*, 1982) or EF-hand (Kretsinger, 1987) are exceptional cases. A short sequence search is limited to the case that there exist an obvious sequence motif (Hofmann *et al.*, 1999). Our finding that module is a unit of function and evolution opens a new way of function assignment and/or prediction. The first trial was described in Chapter IV.

The whole ideas developed in this final chapter are based on a database of module classification. A classification of modules based on their three-dimensional structures and functions will facilitate newly identifying module shufflings, comparing DNA-module and RNA-module interactions, predicting interaction sites of transcription factors and DNA manipulation enzymes and predicting functions of putative protein emerging out of total genome sequencing. The future direction of this work is set to the classification.

Acknowledgements

I will express my appreciation to Dr. Mitiko Gō for her instruction, encouragement, support and patience to my pace of work. In every aspect of the research and study, she is the role model of mine. The same amount of thanks goes to Dr. Tosiyuki Noguti for his support and discussions. Details in mathematical procedure were well adjusted at the discussions with him. Involvement into a theoretical biology was rather accidental to me. An encounter to Dr. Nobuhiko Saito had a huge impact on me. He had changed my perspective toward biology and made me step into the field of theoretical biology. An impressive words out of his mouth was, "a theoretical biologist is nothing but a molecular detective." He once told me in 1989 that the coming research on protein structure should be focused on DNA-protein interactions. Unintentionally, I have followed his visions. I should not forget to thank Dr. Jiri Strakota, his colleagues and staffs of computer centre at high energy physics laboratory in Prague. He did tell me how to program and how to cope with old computer system in a modern way. All of my present skill and techniques on computing derives from his lessons. The most important lesson was, "young man always blames the computer system, but 99.9% you are wrong." The works developed here are, of course, supported by all the staff in the department who tell me the detail of biological system in patience, students in the laboratory who endure my unlimited use of computers, and librarians and managerial staff of the department who kindly accepted my limitless and timeless requests.

Bibliography

Aggarwal A.K., Rodgers D.W., Drottar M., Ptashne M. and Harrison S.C. (1988) Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242:** 899-907.

Albright R.A., Mossing M.C. and Matthews B.W. (1998) Crystal structure of an engineered Cro monomer bound nonspecifically to DNA: Possible implications for nonspecific binding by the wild-type protein. *Prot. Sci.* **7**: 1485-1494.

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

Anderson W.F., Ohlendorf D.H., Takeda Y. and Matthews B.W. (1981) Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature* **290:** 754-758.

Assa-Munt N., Mortishire-Smith R.J., Aurora R., Herr W. and Wright, P.E. (1993) The solution structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage λ repressor DNA-binding domain. *Cell* **73:** 193-205.

Bajaj M. and Blundell T. (1984) Evolution and the tertiary structure of proteins. *Annu. Rev. Biophys. Bioeng.* **13:** 453-492.

Baker E.N. and Hubbard R.E. (1984) Hydrogen bonding in globular proteins. *Prog. Biophys. Molec. Biol.* **44:** 97-179.

Baumeister R., Helbl V. and Hillen W. (1992) Contacts between Tet repressor and tet operator revealed by new recognition specificities of single amino acid replacement mutants. *J. Mol. Biol.* **226:** 1257-1270.

Beese L.S., Friedman J.M. and Steitz T.A. (1993) Crystal structures of the Klenow fragment of DNA polymerase I complexed with deoxynucleoside triphosphate and pyrophosphate. *Biochemistry* **32:** 14095-14101.

Berget S.M., Moore C. and Sharp P.A. (1977) Spliced segements at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* **74:** 3171-3175.

Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer Jr E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T. and Tasumi M. (1977) The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112:** 535-542.

Billeter M., Qian Y.Q., Otting G., Muller M., Gehring W. and Wüthrich K. (1993)

Determination of the nuclear magnetic resonance solution structure of an antennapedia homeodomain-DNA complex. *J. Mol. Biol.* **234:** 1084-1093.

Blake C.C.F. (1978) Do genes-in-pieces imply proteins-in-pieces? *Nature* **273:** 267-268.

Blattner F.R., Plunkett G. III, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J. Mau B. and Shao Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453-1474.

Brennan R.G. and Matthews B.W. (1989) The helix-turn-helix DNA binding motif. *J. Biol. Chem.* **264:** 1903-1906.

Brennan R.G., Roderick S.L., Takeda Y. and Matthews B.W. (1990) Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex. *Proc. Natl. Acad. Sci. USA* **87:** 8165-8169.

Cantor C.R. and Jukes T.H. (1966) The repetition of homologous sequences in the polypeptide chains of certain cytochromes and globins. *Proc. Natl. Acad. Sci. USA* **56**:177-184.

Cheng X., Balendiran K., Schildkraut I. and Anderson J.E. (1994) Structure of PvuII endonuclease with cognate DNA. *EMBO J.* **13:** 3927-3935.

Cho Y., Gorina S,. Jeffrey P.D. and Pavletich N.P. (1994) Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations. *Science* **265:** 346-355.

Choo Y. and Klug A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Strct. Biol.* **7:** 117-125.

Chothia C. and Lesk A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823-826.

Chow L.T., Gelinas R.E., Broker T.R. and Roberts R.J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12:** 1-8.

Christy B.A., Lau L.F. and Nathans D. (1988) A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with "zinc finger" sequences. *Proc. Natl. Acad. Sci. USA* **85:** 7857-7861.

Chuprina V.P., Rullmann J.A., Lamerichs R.M., van Boom J.H., Boelens R. and Kaptein R. (1993) Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J. Mol. Biol.* **234:** 446-462.

Chyan Y.J., Ackerman S., Shepherd N.S., McBride O.W., Widen S.G., Wilson S.H. and Wood T.G. (1994) The human DNA polymerase beta gene structure. Evidence of alternative splicing in gene expression. *Nucl. Acids Res.* **22:** 2719-2725.

Clarke N.D., Beamer L.J., Goldberg H.R., Berkower C. and Pabo C.O. (1991) The DNA binding arm of lambda repressor: Critical contact from a flexible region. *Science* **254:** 267-270.

Connolly M.L. (1983) Analytical molecular surface calculation. *J. Appl. Cryst.* **16:** 548-558.

Darnell J.E. and Doolittle W.F. (1986) Speculations on the early course of evolution. *Proc. Natl. Acad. Sci. USA* **83:** 1271-1275.

Dayhoff M.O., Schwartz R.M. and Orcutt B.C. (1978) A model of evolutionary change in proteins, In: Atlas of protein sequence and structure. (ed. Dayhoff, M.O.) National Biomedical Research Foundation, Washington D.C., pp. 345-352.

Doherty A.J., Serpell L.C. and Ponting C.P. (1996) The helix-hairpin-helix DNA-binding motif: A structural basis for non-sequence-specific recognition of DNA. *Nucl. Acids Res.* **24:** 2488-2497.

Doolittle W.F. (1978) Genes in pieces: Were they ever together? *Nature* **272:** 581-582.

Doolittle R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64:** 287-314.

Dorit R.L., Schoenback L. and Gilbert W. (1990) How big is the universe of exons? *Science* **250:** 1377-1382.

Dorit R.L. and Gilbert W. (1991) The limited universe of exons. *Curr. Opin. Struct. Biol.* **1:** 973-977.

Dreisenkelmann B. (1994) Translocation of DNA across bacterial membranes. *Microbiol. Rev.*, **58:**293-316.

Dubnau J. and Struhl G. (1996) RNA recognition and translational regulation by a homeodomain protein. *Nature* **379:** 694-699.

Eck R.V. and Dayhoff M.O. (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152:** 363-366.

Ellenberger T.E., Brandl C.J., Struhl K. and Harrison S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: Crystal structure of the protein-DNA complex. *Cell* **71:** 1223-1237.

Fairall L., Schwabe J.W., Chapman L., Finch J.T. and Rhodes D. (1993) The crystal structure

of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366:** 483-487.

Feng J.A., Johnson R.C. and Dickerson R.E. (1994) Hin recombinase bound to DNA: The origin of specificity in major and minor groove interactions. *Science* **263:** 348-355.

Finney M., Ruvkun G. and Horvitz H.R. (1988) The *C. elegans* cell lineage and differentiation gene *unc-86* encodes a protein with a homeodomain and extended similarity to transcription factors. *Cell* **55:** 757-769.

Fukami-Kobayashi K. (1994) Estimation of evolutionary distance between distantly related sequences of amino acids, taking account of patterns of amino acid replacement. *Mol. Biol. Evol.* **11:** 99-105.

Gehring W.J., Affolter M. and Bürglin T. (1994) Homeodomain proteins. *Annu. Rev. Biochem.* **63:** 487-526.

Ghosh G., van Duyne G., Ghosh S. and Sigler P.B. (1995) Structure of NF-κ B p50 homodimer bound to a kappa B site. *Nature* **373:** 303-310.

Gilbert W. (1978) Why genes in pieces? *Nautre* **271:** 501.

Gilbert W., Marchionni M. and McKnight G. (1986) On the antiquity of introns. *Cell* **46:**151-154.

Gilbert W. and Glynias M. (1993) On the ancient nature of introns. *Gene* **135:** 137-144.

Glover J.N. and Harrison S.C. (1995) Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373:** 257-261.

Gō M. (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291:** 90-92.

Gō M. (1983) Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* **80:** 1964-1968.

Gō M. (1985) Protein structures and split genes. *Adv. Biophys.* **19:** 91-131.

Gō M. and Miyazawa S. (1980) Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. *Int. J. Peptide Protein Res.* **15**: 211-224.

Gō M. and Nosaka M. (1987) Protein architecture and the origin of intron. *Cold Spring Harbor Symp. Quat. Biol.* **52:** 915-924.

Gō M and Noguti T. (1995) Putative origin of introns deduced from protein anatomy. In: Tracing biological evolution in protein and gene structures. (eds. Gō M. and Schimmel P.) Elsevier, Amsterdam pp. 229-235.

Griffith F. (1928) The significance of pneumococcal types. *J. Hyg.* **27:** 113-159.

Grigorieva G. and Shestakov S. (1982) Transformation in the cyanobacterium *Synechocystis* sp. 6803. *FEMS Microbiol. Lett.***13:** 367-370.

Hahn J., Inamine G., Kozlov Y. and Dubnau D. (1993) Characterization of comE, a late competence operon of *Bacillus subtilis* required for the binding and uptake of transforming DNA. *Mol. Microbiol.* **10:** 99-111.

Hargreaves D., Rice D.W., Sedelnikova S.E., Artymiuk P.J., Lloyd R.G. and Rafferty J.B. (1998) Crystal structure of *E.coli* RuvA with bound DNA Holliday junction at 6Å resolution. *Nature Str. Biol.* **5:** 441-446.

Harrison S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature* **353:** 715-719.

Harrison S.C. and Aggarwal A.K. (1990) DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **59:** 933-969.

Hecht M.H., Sturtevant J.M. and Sauer R.T. (1986) Stabilization of lambda repressor against thermal denaturation by site-directed Gly-Ala changes in alpha-helix 3. *Proteins* **1**: 43-46.

Hegde R.S., Grossman S.R., Laimins L.A. and Sigler P.B. (1992) Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **359:** 505-512.

Hinck A.P., Markus M.A., Huang S., Grzesiek S., Kustonovich I., Draper D.E. and Torchia D.A. (1997) The RNA binding domain of ribosomal protein L11: Three-dimensional structure of the RNA-bound form of the protein and its interaction with 23S rRNA. *J. Mol. Biol.* **274:** 101-113.

Hofmann K., Bucher P., Falquet, L. and Bairoch, A (1999) The PROSITE database, its status in 1999. *Nuc. Acids Res.* **27:** 215-219.

Hol W.G.J., van Duijnen P.T. and Berendsen H.J.C. (1978) The α-helix dipole and the properties of proteins. *Nature* **273:** 443-446.

Houbaviy H.B., Usheva A., Shenk T. and Burley S.K. (1996) Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl. Acad. Sci. USA* **93:**13577-13582.

Ikura T., Gō N., Kohda D., Inagaki F., Yanagawa H., Kawabata M., Kawabata S., Iwanaga S., Noguti T. and Gō M. (1993) Secondary structural features of modules M2 and M3 of barnase in solution by NMR experiment and distance geometry calculation. *Proteins:*

Strct. Funct. Genet. **16:** 341-356.

Inaba K., Ishimori K., Imai K. and Morishima I. (1998) Structural and functional effects of pseudo-module substitution in hemoglobin subunits. New structural and functional units in globin structure. *J. Biol. Chem.* **273:** 8080-8087.

Inamine G.S. and Dubnau D. (1995) ComEA, a *Bacillus subtilis* integral membrane protein required for genetic transformation, is needed for both DNA binding and transport. *J. Bacteriol.* **177:** 3045-3051.

Jay D.G. and Gilbert W. (1987) Basic protein enhances the incorporation of DNA into lipid vesicles: Model for the formation of primordial cells. *Proc. Natl. Acad. Sci. USA.* **84:** 1978-1980.

Jordan S.R. and Pabo C.O. (1988) Structure of the lambda complex at 2.5 Å resolution: Details of the repressor-operator interactions. *Science* **242:** 893-899.

Joyce G.F (1987) Nonenzymatic template-directed synthesis of informtaional macromolecules. *Cold Spring Harbor Symp. Quat. Biol.* **52:** 41-51.

Joyce G.F. and Orgel L.E. (1993) Prospects for understanding the origin of the RNA world. In: The RNA world. (eds. Gesteland R.F and Atkins J.F.) Cold Spring Harbor Laboratory Press. pp. 1-25.

Joyce C.M. and Steitz T.A. (1994) Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* **63:** 777-822.

Kamada K., Horiuchi T., Ohsumi K., Shimamoto N. and Morikawa K. (1996) Structure of a replication-terminator protein complexed with DNA. *Nature* **383:** 598-603.

Kaneko T., Sato S., Kotani H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Hirosawa M., Sugiura M., Sasamoto S., Kimura T., Hosouchi, T., Matsuno A., Muraki A., Nakazaki N., Naruo K., Okumura, S., Shimpo S., Takeuchi C., Wada T., Watanabe A., Yamada M., Yasuda M. and Tabata S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109-136.

Kaneko T. and Tabata, S. (1997) Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. *Plant Cell Physiol.* **38:** 1171-1176.

Kelley R.L. and Yanofsky C. (1985) Mutational studies with the *trp* repressor of *Escherichia coli* support the helix-turn-helix model of repressor recognition of operator DNA. *Proc.*

Natl. Acad. Sci. USA **82:** 483-487.

Kim Y., Geiger J.H., Hahn S. and Sigler P.B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365:** 512-520.

Kim J.L., Nikolov D.B. and Burley S.K. (1993) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365:** 520-527.

Kisker C., Hinrichs W., Tovar K., Hillen W. and Saenger W. (1995) The complex formed between tet repressor and tetracycline-Mg^{2+} reveals mechanism of antibiotic resistance. *J. Mol. Biol.* **247:** 260-280.

Kissinger C.R., Liu B.S., Martin-Blanco E., Kornberg T.B. and Pabo C.O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell* **63:** 579-590.

Klemm J.D., Rould M.A., Aurora R., Herr W. and Pabo C.O. (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77:** 21-32.

Klimasauskas S., Kumar S., Roberts R.J. and Cheng X. (1994) HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **76:** 357-369.

Koonin E.V. (1996) A duplicated catalytic motif in a new superfamily of phosphohydrolases and phospholipid synthases that includes poxvirus envelope proteins. *Trends Biochem. Sci.* **21:** 242-243.

Koonin E.V. and Galperin M.Y. (1997) Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr. Opin Genet. Dev.* **7:** 757-763.

Kostrewa D. and Winkler F.K. (1995) Mg^{2+} binding to the active site of EcoRV endonuclease: A crystallographic study of complexes with substrate and product DNA at 2 Å resolution. *Biochemistry* **34:** 683-696.

Kretsinger R.H. (1987) Calcium coordination and the calmodulin fold: Divergent versus convergent evolution. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 499-510.

Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessières P., Bolotin A., Borchert S., Borriss R., Boursier L., Brans A., Braun M., Brignell S.C., Bron S., Brouillet S., Bruschi C.V., Caldwell B., Capuano V., Carter N.M., Choi S.-K., Codani J.-J., Connerton I.F., Cummings N.J., Daniel R.A., Denizot F., Devine K.M., Düsterhöft A., Ehrlich S.D., Emmerson P.T., Entian K.D., Errington J., Fabret C., Ferrari E., Foulger D., Fritz C., Fujita M., Fujita Y., Fuma S., Galizzi A.,

Galleron N., Ghim S.-Y., Glaser P., Goffeau A., Golightly E.J. Grandi G., Guiseppi G., Guy B.J., Haga K., Haiech J., Harwood C.R., Hénaut A., Hilbert H., Holsappel S., Hosono S., Hullo M.-F., Itaya M., Jones L., Joris B., Karamata D., Kasahara Y., Klaerr-Blanchard M., Klein C., Kobayashi Y., Koetter P., Koningstein G., Krogh S., Kumano M., Kurita K., Lapidus A., Lardinois S., Lauber J., Lazarevic V., Lee S.-M., Levine A., Liu H., Masuda S., Mauël C., Médigue C., Medina N., Mellado R.P., Mizuno M., Moestl D., Nakai S., Noback M., Noone D., O'Reilly M., Ogawa K., Ogiwara A., Oudega, B., Park S.-H., Parro V., Pohl T.M., Protetelle D., Porwollik S., Prescott A.M., Presecan E., Pujic P., Purnelle B., Rapoport G., Rey M., Reynolds S., Rieger M., Rivolta C., Rocha E., Roche B., Rose M., Sadaie Y., Sato T., Scanlan E., Schleich S., Schroeter R., Scoffone F., Sekiguchi J., Sekowska A., Seror S.J., Serror P., Shin B.-S., Soldo B., Sorokin A., Tacconi E., Takagi T., Takahashi H., Takemaru K., Takeuchi M., Tamakoshi A., Tanaka T., Terpstra P., Tognoni A., Tosato V., Uchiyama S., Vandenbol M., Vannier F., Vassarotti A., Viari A., Wambutt R., Wedler E., Wedler H., Weitzenegger T., Winters P., Winpat A., Yamamoto H., Yamane K., Yasumoto K., Yata K., Yoshida K., Yoshikawa, H.-F., Zumstein E., Yoshikawa H. and Danchi A. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249-256.

Kyte J. and Doolittle R.F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157:** 105-132.

Lahm A. and Suck D. (1991) DNase I-induced DNA conformation. 2 Å structure of a DNase I-octamer complex. *J. Mo.l Biol.* **222:** 645-667.

Lewis M., Chang G., Horton N.C., Kercher M.A., Pace H.C., Schumacher M.A., Brennan R.G. and Lu P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271:** 1247-1254.

Li T., Stark M.R., Johnson A.D. and Wolberger C. (1995) Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science* **270:** 262-269.

Li Y., Itadani H., Sugita M. and Sugiura M. (1992) cDNA cloning and sequencing of tobacco chloroplast ribosomal protein L12. *FEBS Lett.* **300**: 199-202.

Long M., de Souza S.J., Rosenberg C. and Gilbert W. (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl. Acad. Sci. USA* **93:** 7727-7731.

Love J.J., Li X., Case D.A., Giese K., Grosschedl R. and Wright P.E. (1995) Structural basis

for DNA bending by the architectural transcription factor LEF-1. *Nature* **376:** 791-795.

Luisi, B. (1995) DNA-protein interaction at high resolution. In: DNA-protein: Structural interactions. (ed. Lilley, D.M.J) IRL Press Oxford University Press, Oxford, pp. 1-48.

Luisi B.F., Xu W.X., Otwinowski Z., Freedman L.P., Yamamoto K.R. and Sigler P.B. (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **1352:** 497-505.

Ma P.C., Rould M.A., Weintraub H. and Pabo C.O. (1994) Crystal structure of MyoD bHLH domain-DNA complex: Perspectives on DNA recognition and implications for transcriptional activation. *Cell* **77:** 451-459.

Mandel-Gutfreund Y., Schueler O. and Margalit H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: In search of common principles. *J. Mol. Biol.* **253:** 370-382.

Marmorstein R., Carey M., Ptashne M. and Harrison S.C. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356:** 408-414.

Marmorstein R. and Harrison S.C. (1994) Crystal structure of a PPR1-DNA complex: DNA recognition by proteins containing a Zn_2Cys_6 binuclear cluster. *Genes Dev.* **8:** 2504-2512.

Matsuo K., Clay O., Kunzler P., Georgiev O., Urbanek P. and Schaffner W. (1994) Short introns interrupting the Oct-2 POU domain may prevent recombination between POU family genes without interfering with potential POU domain 'shuffling' in evolution. *Biol. Chem. Hoppe-Seyler* **375:** 675-683.

Mattaj I.W. and Nagai K. (1995) Recruiting proteins to the RNA world. *Nature Str. Biol.* **2:**518-522.

Mayo S.L., Olafson B.D. and Goddard W.A.,III (1990) DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94**: 8897-8909.

McClarin J.A., Frederick C.A., Wang B.C., Greene P., Boyer H.W., Grable J. and Rosenberg J.M. (1986) Structure of the DNA-Eco RI endonuclease recognition complex at 3 Å resolution. *Science* **234:** 1526-1541.

McLachlan A.D. (1971) Tests for comparing related amino-acid sequences. Cytochorme *c* and cytochrome c_{551}. *J. Mol. Biol.* **61:** 409-424.

McLachlan A.D. (1972) Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **64:**417-437.

McLachlan A.D. (1979) Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128:** 49-79.

McLachlan A.D. (1987) Gene duplication and the origin of repetitive protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 411-420.

Mondragon A. and Harrison S.C. (1991) The phage 434 Cro/OR1 complex at 2.5 Å resolution. *J. Mol. Biol.* **219:** 321-334.

Myers L.C. and Verdine G.L. (1994) DNA repair proteins. *Curr. Opin. Str. Biol.* **4:**51-59.

Nagadoi A., Morikawa S., Nakamura H., Enari M., Kobayashi K., Yamamoto H., Sampei. G., Mizobuchi K., Schumacer M.A., Brennan R.G. and Nishimura Y. (1995) Structural comparison of the free and DNA-bound forms of the purine repressor DNA-binding domain. *Structure* **3:** 1217-1224.

Neri D., Billeter M., Wider G. and Wüthrich K. (1992) NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science* **257:** 1559-1563.

Nikolov D.B., Chen H,. Halay E.D., Usheva A.A., Hisatake K., Lee D.K., Roeder R.G. and Burley S.K. (1995) Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **377:** 119-128.

Nishino T., Ariyoshi M., Iwasaki H., Sninagawa H. and Morikawa K. (1998) Functional analyses of the domain structure in the Holliday junction binding protein RuvA.*Structure* **6:** 11-21.

Noguti T., Sakakibara H. and Gō M. (1993) Localization of hydrogen-bonds within modules in barnase. *Proteins: Strct. Funct. Genet.* **16:** 357-363.

Nureki O., Vassylyev D.G., Katayanagi K., Shimizu T., Sekine S., Kigawa T., Miyazawa T., Yokoyama S. and Morikawa K. (1995) Architectures of class-defining and specific domains of glutamyl-tRNA synthetase. *Science* **267:** 1958-1965.

Ogata K., Morikawa S., Nakamura H., Sekikawa A., Inoue T., Kanai H., Sarai A., Ishii S. and Nishimura Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79:** 639-648.

Ohlendorf D.H., Anderson W.F. and Matthews B.W. (1983) Many gene-regulatory proteins appear to have a similar alpha-helical fold that binds DNA and evolved from a common precursor. *J. Mol. Evol.* **9:** 109-114.

Ohno S. (1984) Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. *J. Mol. Evol.* **20**: 313-321.

Ohno S. (1987) Evolution from primordial oligomeric repeats to modern coding sequences. *J. Mol. Evol.* **25:** 325-329.

Omichinski J.G., Clore G.M., Schaad O., Felsenfeld G., Trainor C., Appella E., Stahl S.J. and Gronenborn A.M.(1993) NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science* **261:** 438-446.

Okazawa H., Okamoto K., Ishino F., Ishino-Kaneko T., Takeda S., Toyoda Y., Muramatsu M. and Hamada H. (1991) The oct3 gene, a gene for an embryonic transcription factor, is controlled by a retinoic acid repressible enhancer. *EMBO J.* **10:** 2997-3005.

Otting G., Qian Y.Q., Billeter M., Muller M., Affolter M., Gehring W.J. and Wüthrich K. (1990) Protein--DNA contacts in the structure of a homeodomain-DNA complex determined by nuclear magnetic resonance spectroscopy in solution. *EMBO J.* **9:** 3085-3092.

Otwinowski Z., Schevitz R.W., Zhang R.G., Lawson C.L., Joachimiak A., Marmorstein R.Q., Luisi B.F. and Sigler P.B. (1988) Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature* **335:** 321-329.

Pabo C.O. and Sauer R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61:** 1053-1095.

Patthy L. (1991) Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1:** 351-361.

Patthy L. (1996) Exon shuffling and other ways of module exchange. *Matrix Biol.* **15:** 301-310.

Pavletich N.P. and Pabo C.O. (1991) Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252:** 809-817.

Pellegrini L., Tan S. and Richmond T.J. (1995) Structure of serum response factor core bound to DNA. *Nature* **376:** 490-498.

Pelletier H., Sawaya M.R., Kumar A., Wilson S.H. and Kraut J. (1994) Structures of ternary complexes of rat DNA polymerase β, a DNA template-primer, and ddCTP. *Science* **264:** 1891-1903.

Pelletier H., Sawaya M.R., Wolfle W., Wilson H. and Kraut J. (1996) Crystal structures of human DNA polymerase β complexed with DNA: Implications for catalytic mechanism, processivity, and fidelity. *Biochemistry* **35:** 12742-12761.

Pohlman R.F., Liu F., Wang L., Moré M.I. and Winans S.C. (1993) Genetic and biochemical

analysis of an endonuclease encoded by the IncN plasmid pKM101. *Nuc. Acids Res.* **21:** 4867-4872.

Ponting C.P. and Kerr I.D. (1996) A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: Identification of duplicated repeats and potential active site residues. *Prot. Sci.* **5:** 914-922.

Porter R.D. (1986) Transformation in cyanobacteria. *CRC Crit. Rev. Microbiol.* **13:** 111-131.

Rafferty J.B., Somers W.S., Saint-Girons I. and Phillips S.E. (1989) Three-dimensional crystal structures of *Escherichia coli* met repressor with and without corepressor. *Nature* **341:** 705-710.

Ramakrishnan V., Finch J.T., Graziano V., Lee P.L. and Sweet R.M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362:** 219-223.

Rastinejad F., Perlmann T., Evans R.M. and Sigler P.B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* **375:** 203-211.

Raumann B.E., Brown B.M. and Sauer R.T. (1994a) Major groove DNA recognition by β-sheets: The ribbon-helix-helix family of gene regulatory proteins. *Curr. Opin. Strct. Biol.* **4:** 36-43.

Raumann B.E., Rould M.A., Pabo C.O. and Sauer R.T. (1994b) DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature* **367:** 754-757.

Remington S.J. and Matthews B.W. (1980) A systematic approach to the comparison of protein structures. *J. Mol. Biol.* **140**: 77-99.

Rice P.A. (1997) Making DNA do a U-turn; IHF and related proteins. *Curr. Opin. Strct. Biol.* **7:** 86-93.

Rice P.A. and Steitz T.A. (1989) Ribosomal protein L7/L12 has a helix-turn-helix motif similar to that found in DNA-binding regulatory proteins. *Nucleic Acids Res.* **17**: 3757-3762.

Rice P.A., Yang S., Mizuuchi K. and Nash H.A. (1996) Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87:** 1295-1306.

Richardson J.S. and Richardson D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240:** 1648-1652.

Richardson J.S. and Richardson D.C. (1988) Helix lap-joints as ion-binding sites: DNA-binding motifs and Ca-binding "EF hands" are related by

charge and sequence reversal. *Proteins* **4**: 229-239.

Riley L.K., Morrow J.K., Danton M.J. and Coleman M.S. (1988) Human terminal deoxyribonucleotidyltransferase: Molecular cloning and structural analysis of the gene and 5' flanking region. *Proc. Natl. Acad. Sci. USA* **85:** 2489-2493.

Saitou N. and Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.

Sauer R.T., Jordan S.R. and Pabo C.O. (1990) λ repressor: A model system for understanding protein-DNA interactions and protein stability. *Adv. Protein Chem.* **40:** 1-61.

Sawaya M.R., Pelletier H., Kumar A., Wilson S.H. and Kraut J. (1994) Crystal structure of rat DNA polymerase β: evidence for a common polymerase mechanism. *Science* **264:** 1930-1935.

Sawaya M.R., Prasad R., Wilson S.H., Kraut J. and Pelletier H. (1997) Crystal structures of human DNA polymerase β complexed with gapped and nicked DNA: Evidence for an induced fit mechanism. *Biochemistry* **36:** 11205-11215.

Schevitz R.W., Otwinowski Z., Joachimiak A., Lawson C.L. and Sigler P.B. (1985) The three-dimensional structure of *trp* repressor. *Nature* **317:** 782-786.

Schleif R. (1988) DNA binding by proteins. *Science* **241:** 1182-1187.

Schultz S.C., Shields G.C. and Steitz T.A. (1991) Crystal structure of a CAP-DNA complex: The DNA is bent by 90 degrees. *Science* **253:** 1001-1007.

Schumacher M.A., Choi K.Y., Zalkin H. and Brennan R.G. (1994) Crystal structure of LacI member, PurR, bound to DNA: Minor groove binding by alpha helices. *Science* **266:** 763-770.

Schwabe J.W., Chapman L., Finch J.T. and Rhodes D. (1993) The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: How receptors discriminate between their response elements. *Cell* **75:** 567-578.

Seeberg E., Eide L. and Bjørås M. (1995) The base excision repair pathway. *Trends Biochem. Sci.* **20:** 391-397.

Sharp P. A. (1994) Split genes and RNA splicing. Cell 805-815.

Sigler P.B. (1992) The molecular mechanism of *trp* repression. In: Transcriptional regulation. (eds. McKnight S.L. and Yamamoto K.R.) Cold Spring Harbor Laboratory Press, New York pp 475-499.

Smith E.L., DeLange R.J., Evans W.H., Landon M. and Markland F.S. (1968) Subtilisin

Carlsberg V. The complete sequence; comparison with subtilisin BPN'; evolutionary relationships. *J. Biol. Chem.* **243:** 2184-2191.

Solomon J.M. and Grossman A.D. (1996) Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet.*, **12:** 150-155.

de Souza S.J., Long M., Schoenbach L., Roy S.W. and Gilbert W. (1996) Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA* **93:** 14632-14636.

Steitz T.A. (1990) Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Q. Rev. Biophys.* **23:** 205-280.

Stephens R.S., Kalman S., Lammel C., Fan J., Marathe R., Aravind L., Mitchell W., Olinger L., Tatusov R.L., Zhao Q., Koonin E.V. and Davis R.W. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282:** 754-759.

Strimmer K. and von Haeseler A. (1996) Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13:** 964-969.

Suzuki M. and Brenner S.E. (1995) Classification of multi-helical DNA-binding domains and application to predict the DBD structures of σ factor, LysR, OmpR/PhoB, CENP-B, Rap1, and XylS/Ada/AraC. *FEBS Lett.* **372:** 215-221.

Tainer J.A., Thayer M.M. and Cunningham R.P. (1995) DNA repair proteins. *Curr. Opin. Str. Biol.* **5:** 20-26.

Takahashi K., Oohashi M., Noguti T. and Gō M. (1997) Mechanical stability of compact modules of barnase, *FEBS Lett.* **405:** 47-54.

Takeda Y., Ohlendorf D.H., Anderson W.F. and Matthews B.W. (1983) DNA-binding proteins. *Science* **221:** 1020-1026.

Tan S. and Richmond T.J. (1998) Eukaryotic transcription factors. *Curr. Opin. Strct. Biol.* **8:** 41-48.

Tan S., Hunziker Y., Sargent D.F. and Richmond T.J. (1996) Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* **381:** 127-131.

Tanaka I., Appelt K., Dijk J., White S.W. and Wilson K.S. (1984) 3-Å resolution structure of a protein with histone-like properties in prokaryotes. *Nature* **310:** 376-381.

Tasayco M.L. and Carey J. (1992) Ordered self-assembly of polypeptide fragments to form nativelike *trp* repressor. *Science* **255:** 594-597.

Tateno M., Mizutani M., Yura K., Nureki O., Yokoyama S. and Gō M. (1995) Module Structure and Function of Glutamyl-tRNA Synthetase. In: Tracing Biological Evolution in Protein and Gene Structures. (eds. Gō M. and Schimmel P.) Elsevier, Amsterdam pp. 53-63.

Thayer M.M., Ahern H., Xing D., Cunningham R. and Tainer J.A. (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J.* **14:** 4108-4120.

Theill L.E., Hattori K., Lazzaro D., Castrillo J-L. and Karin M. (1992) Differential splicing of the *GHF1* primary transcript gives rise to two functionally distinct homeodomain proteins. *EMBO J.* **11:** 2261-2269

Tittiger C., Whyard S. and Walker V.K. (1993) A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*. *Nature* **361:** 470-472.

Vassylyev D.G., Kashiwagi T., Mikami Y., Ariyoshi M., Iwai S., Ohtsuka E., Morikawa K. (1995) Atomic model of a pyrimidine dimer excision repair enzyme complexed with a DNA substrate: Structural basis for damaged DNA recognition. *Cell* **83:** 773-782.

Vassylyev D.G. & Morikawa K. (1997) DNA-repair enzymes. *Curr. Opin. Str. Biol.* **7:** 103-109.

Wah D.A., Hirsch J.A., Dorner L.F., Schildkraut I. and Aggarwal A.K. (1997) Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388:** 97-100.

Wakasugi K., Ishimori K., Imai K., Wada Y. and Morishima I. (1994) "Module" substitution in hemoglobin subunits. *J. Biol. Chem.* **269:** 18750-18756.

Walker J.E., Saraste M., Runswick M.J. and Gay N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1:** 945-951.

Warwicker J. and Watson H.C. (1982) Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **157:** 671-679.

Werner M.H., Clore M., Fisher C.L., Fisher R.J., Trinh L., Shiloach J. and Gronenborn A.M. (1995a) The solution structure of the human ETS1-DNA complex reveals a novel mode of binding and true side chain intercalation. *Cell* **83:** 761-71.

Werner M.H., Huth J.R., Gronenborn A.M. and Clore G.M. (1995b) Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* **81:** 705-714.

Wetlaufer D.B. (1973) Nucleation, rapid folding and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA* **70:** 697-701.

White S.H. (1994) Global statistics of protein sequences: Implications for the origin, evolution, and prediction of structure. *Annu. Rev. Biophys. Biomol. Struct.* **23:** 407-439.

Wilson K.P., Shewchuk L.M., Brennan R.G., Otsuka A.J. and Matthews B.W. (1992) *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. *Proc. Natl. Acad. Sci. USA* **89:** 9257-9261.

Wilson R., Ainscough R., Anderson K., Baynes C., Berks M., Bonfield J., Burton J., Connell M., Copsey T., Cooper J., Coulson A., Craxton M., Dear S., Du Z., Durbin R., Favello A., Fraser A., Fulton L., Gardner A., Green P., Hawkins T., Hillier L., Jier M., Johnston L., Jones M., Kershaw J., Percy C., Rifken L., Roopra A., Saunders D., Shownkeen R., Sims M., Smaldon N., Smith A., Smith M., Sonnhammer E., Staden R., Sulston J., Thierry-Mieg J., Thomas K., Vaudin M., Vaughan K., Waterston R., Watson A., Weinstock L., Wilkinson-Sproat J. and Wohldman P. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368:** 32-38.

Wintjens R.T., Rooman M.J. and Wodak S.J. (1996) Automatic classification and analysis of αα-turn motifs in proteins. *J. Mol. Biol.* **255:** 235-253.

Wolberger C., Vershon A.K., Liu B., Johnson A.D. and Pabo C.O. (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67:** 517-528.

Wong E.A., Silsby J.L. and Halawani M.E.E. (1992) Complementary DNA cloning and expression of Pit-1/GHF-1 from the domestic turkey. *DNA Cell Biol.* **11:** 651-660.

Xu W., Rould M.A., Jun S., Desplan C. and Pabo C.O. (1995) Crystal structure of a paired domain-DNA complex at 2.5Å resolution reveals structural basis for Pax developmental mutations. *Cell* **80:** 639-650.

Yuan H.S., Finkel S.E., Feng J.-A., Kaczor-Grzeskowiak M., Johnson R.C. and Dickerson R.E. (1991) The molecular structure of wild-type and a mutant Fis protein: Relationship between mutational changes and recombinational enhancer function or DNA binding. *Proc. Natl Acad. Sci. USA* **88:** 9558-9562.

Yanagawa H., Yoshida K., Torigoe C., Park J-S., Sato K., Shirai T. and Gō M. (1993) Protein anatomy: functional roles of barnase module. *J. Biol. Chem.* **268:** 5861-5865.

Yaoi T., Miyazaki K., Oshima T., Komukai Y. and Gō M. (1996) Conversion of the

coenzyme specificity of isocitrate dehydrogenase by module replacement. *J. Biochem. (Tokyo)* **119:** 1014-1018.

Yura K., Tomoda S. and Gō M. (1993) Repeat of a helix-turn-helix module in DNA binding proteins. *Protein Engng*, **6**:621-628.

Yura K. and Gō M. (1995) Helix-turn-helix module distribution and module shuffling. In: Tracing Biological Evolution in Protein and Gene Structures. (eds. Gō M. and Schimmel P.) Elsevier, Amsterdam pp. 187-195.

Yura K. and Gō M. (1997) The homeodomain-like putative product of plastid genome: A possible role in plastid differentiation. *Res. Comm. Biochem. Cell Mol. Biol.* **1:** 79-81.

Yura K., Shionyu M., Kawatani K. and Gō, M. (1999a) Repetitive use of a phosphate-binding module in DNA polymerase β, Oct-1 POU domain and phage repressors. *Cell. Mol. Life Sci.* in press.

Yura K., Toh H. and Gō M. (1999b) Putative Mechanism of Genetic Transformation as Deduced from Genome Data. *DNA Res.* in press.

Zuckerkandl E. (1994) Molecular pathways to parallel evolution: I. Gene nexuses and their mophological correlates. *J. Mol. Evol.* **39:** 661-678.

副論文

1. Repetitive use of a phosphate-binding module in DNA polymerase β, Oct-1 POU domain and phage repressors.
   Yura, K., Shionyu, M., Kawatani, K., and M. Gō, M.
   Cell. Mol. Life Sci. in press
   （DNAポリメラーゼβ、Oct-1　POUドメイン及びファージのリプレッサーに繰り返し用いられているリン酸基結合モジュール）

2. Putative Mechanism of Natural Transformation as Deduced from Genome Data.
   Yura, K., Toh, H., and Gō, M.
   DNA Res. in press
   （ゲノム情報から予想される形質転換のメカニズム）

3. Helix-turn-helix module distribution and module shuffling.
   Yura, K., and Gō, M.
   In: Tracing Biological Evolution in Protein and Gene Structures. (eds. Gō, M. and Schimmel, P.), Elsevier, Amsterdam, 187-195, 1995.
   （ヘリックス・ターン・ヘリックスモジュールの分配とモジュール・シャッフリング）

4. Repeat of a helix-turn-helix module in DNA binding proteins.
   Yura, K., Tomoda, S. and Gō, M.
   Protein Engng, **6:** 621-628, 1993.
   （DNA結合タンパク質におけるヘリックス・ターン・ヘリックスモジュールの繰り返し）